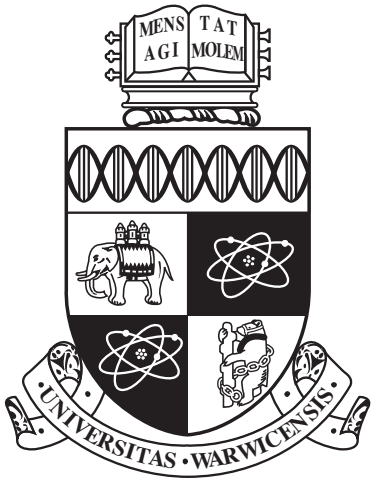


Studying Effective Brain Connectivity Using Multiregression Dynamic Models

by

Lilia Carolina Carneiro da Costa

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

The Department of Statistics

November 2014

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgments **v**

Declarations **vi**

Abstract **vii**

Abbreviations **viii**

Chapter 1 Introduction **1**

1.1 Thesis Outline 2

1.2 Resting-state 4

1.3 Functional Magnetic Resonance Imaging 5

Chapter 2 Brain Connectivity **9**

2.1 Structural, Functional and Effective Connectivity 10

2.2 Some Definitions of Graphical Modelling 11

2.3 Some Methods for Discovering Connectivity 13

2.4 Bayesian Network 23

2.4.1 What is a Bayesian Network? 23

2.4.2 The Joint Probability Distribution of the BN 24

2.4.3 Dynamic Bayesian Network (DBN) 26

Chapter 3 The Multiregression Dynamic Model **28**

3.1 Introduction 28

3.2 The Linear MDM 30

3.2.1 The Description of the Model 30

3.2.2 The Inferential Process . . . 34
3.2.3 Priors . . . 37
3.2.4 Criteria for Model Selection . . . 40
3.3 The Process of Search Networks Applied to the MDM . . . 40
3.3.1 Scoring the MDM Using an Integer Programming Algorithm . . . 43
3.3.2 Directed Graph Model Search . . . 46
3.3.3 The Running Time of the MDM-IPA and the MDM-DGM . . . 49
3.4 A Comparison with Some Other Methods . . . 49
3.5 Diagnostic Analysis . . . 53
3.5.1 Global Monitor . . . 54
3.5.2 Parent-child Monitor . . . 56
3.5.3 Node Monitor . . . 57

Chapter 4 The Evaluation Methodology 69

4.1 The MDM Assessment . . . 70
4.2 An Application of the MDM-IPA . . . 74
4.2.1 A DCM Synthetic Study . . . 74
4.2.2 An MDM Synthetic Study . . . 79
4.3 The ICA Data Analysis . . . 83
4.4 A 4-node Resting-State FMRI Data Analysis . . . 87
4.5 Discussion . . . 95

Chapter 5 Group analysis using the MDM 98

5.1 Background . . . 98
5.2 The VTS, the IS and the CS Applied to the MDM . . . 101
5.3 Clustering with Pairwise Log Bayes Factor Separation . . . 102
5.4 Comparing Methods Using Synthetic Data . . . 104
5.4.1 Simulating Data . . . 104
5.4.2 The GS Approach . . . 106
5.4.3 Comparing Group Analysis Approaches . . . 108
5.4.4 Comparing Separation Measures . . . 109
5.5 Group-structure using the Real RS fMRI Data . . . 109

5.5.1 VTS, CS and IS Approaches . . . 112
5.5.2 Comparing Sessions . . . 112
5.5.3 The Application of the Group-structure Approach . . . 114
5.5.4 The GS Approach with the MDM-DGM algorithm . . . 115
5.6 Discussion . . . 115

Chapter 6 Estimation of multiple networks **118**
6.1 Introduction . . . 118
6.2 The IEMN and the MEMN . . . 120
6.2.1 The Individual Estimation of Multiple Networks (IEMN) . . . 120
6.2.2 The Marginal Estimation of Multiple Networks (MEMN) . . . 122
6.3 The Application of Multiple Networks . . . 125
6.3.1 Applying the Individual-Structure Approach . . . 126
6.3.2 Comparing the MEMN with the IEMN in Practice . . . 129
6.4 The Joint Estimation of Multiple Networks (JEMN) . . . 140
6.4.1 A Statistical Model for Joint Multi-Subject Analysis . . . 141
6.4.2 The Application of the JEMN into a Real FMRI Data . . . 145
6.5 Discussion . . . 151

Chapter 7 Further Research **154**
7.1 Search methods for the MDM using non-local priors . . . 154
7.2 The Multiregression Dynamic Hierarchical Models . . . 155

References **158**

A Supplemental Material for Chapter 3 **A1**

B Supplemental Material for Chapter 4 **B1**
Appendix B.1: Comparing Markov equivalent DAGs . . . B1
Appendix B.2: Assessment the directionality using the logBF . . . B1

C Supplemental Material for Chapter 5 **C1**

D Supplemental Material for Chapter 6 **D1**
Appendix D.1: Supplementary material . . . D1

Appendix D.2: The use of diagnostics in a high-dimensional fMRI data D4
Appendix D.3: Additional figures D8

Acknowledgments

I would never have been able to finish my thesis without the guidance of my supervisors, encouragement from family and support from friends.

First and foremost, I would like to express my deepest gratitude to God for answering my prayers, giving me strength, courage and patience in the face of numerous obstacles. I would like to thank my research supervisors Professor Tom Nichols and Professor Jim Smith for their aspiring guidance, support, inspiration and invaluably constructive criticism during this work.

The financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Universidade Federal da Bahia (UFBA), both in Brazil, is greatly acknowledged. I am deeply grateful to Professor Dani Gamerman and lectures, researchers and students from UFBA for encouraging me with their best wishes. Especially, Dr. Rosana Castro who played an important role in my professional life and was always there cheering me up. Sadly, she passed away few days before my viva and I dedicate this thesis to her.

Many people have had a direct impact upon this research, including my examiners Dr. Julia Brettschneider and Dr. Catriona Queen, and researchers Dr. James Cussens, Dr. Chris Oates and Dr. Mark Bartlett. I appreciate all their contributions of time and ideas. I must also thank my friends Alexandre, Roberta, Dragana, Brian, Xu, Nat, the Neurostat group and the Brazilian group for their personal and professional support during the time I spent at the university. I would also like to thank Sandro, Milena, Fabio and Monica for their support and friendly advice when my family and I first arrived in the UK.

Most importantly, none of this could have happened without the help of my family. I owe my deepest gratitude to Rogerio, Lucia, Rosana, Diana, Nelson, Emerson, Rosa, as well as my close friend Vitor and other friends who have always supported, encouraged and believed in me. Last but not least, words cannot express how grateful I am to my husband Alberto and my son Davi for changing their lives dramatically to make my dream come true. This thesis stands as a testament to their unconditional love and encouragement.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. Where I have consulted the published work of others, this is always clearly attributed. The work presented (including data analysis) was carried out by the author except in the cases outlined as follows: (i) all pre-processes of real fMRI data, including the ICA procedure described in Section 4.3; (ii) the application of the JEMN into a real data presented in Section 6.4.2 was done by Christopher Oates jointly with the author.

Parts of this thesis have been published or are in submission:

- Costa, L., Smith, J.Q., Nichols, T. (2013) On the selection of Multiregression Dynamic Models of fMRI networked time series. *CRiSM Working Paper, University of Warwick*, 13(6).

- Costa, L., Smith, J.Q., Nichols, T. (2014) Studying the Effective Brain Connectivity using the Multiregression Dynamic Models. Poster session presented at: *Organization for Human Brain Mapping 2014 Annual Meeting*; Jun 8-12; Hamburg; Germany.

- Costa, L., Smith, J.Q., Nichols, T., Cussens, J., Duff, E.P., Makin, T.R. (2014) Searching Multiregression Dynamic Models of Resting-State fMRI Networks using Integer Programming. *Bayesian Analysis* (to appear).

- Oates, C., Costa, L., Nichols, T. (2014) Towards a Multi-Subject Analysis of Neural Connectivity. *Neural Computation* (to appear).

Abstract

A Multiregression Dynamic Model (MDM) is a class of multivariate time series that represents multiple dynamic causal processes in a graphical way. One of the advantages of this class is that, in contrast to many other Dynamic Bayesian Networks, the hypothesised relationships accommodate conditional conjugate inference. We demonstrate for the first time how it is straightforward to search over all possible connectivity networks with dynamically changing intensity of transmission to find the Maximum a Posteriori Probability (MAP) model within this class. This search method is made feasible by using a novel application of the integer programming algorithm. The search over all possible directed (acyclic or cyclic) graphical structures can be made especially fast by utilising the fact that, within this class of models, the joint likelihood factorizes. We proceed to show how diagnostic methods, analogous to those defined for static Bayesian Networks, can be used to suggest embellishment of the model class to extend the process of model selection.

A typical goal of experimental neuroscience is to draw conclusions regarding the causal mechanisms that underpin neural communication. Often the main focus of interest in these experiments includes not only a search for the likely model of a specific individual, but an analysis of shared between-subject effects. Currently, such features are analysed using rather coarse aggregation methods over shared time series. However, here we demonstrate that, using the estimation of multiple causal graphical models and Bayesian hyperclustering techniques, it is possible to use the full machinery of Bayesian methods to formally make inferences in a coherent way which contemplates hypotheses about shared dependences between such populations of subjects. Methods developed here are illustrated using simulated and real resting-state and steady-state task functional Magnetic Resonance Imaging (fMRI) data.

Abbreviations

BDS	Bilinear Dynamic System
BF	Bayes Factor
BIC	Bayes Information Criterion
BN	Bayesian Network
BOLD	Blood-Oxygenation-Level-Dependent
DAG	Directed Acyclic Graph
DBN	Dynamic Bayesian Network
DCG	Directed Cyclic Graph
DCM	Dynamic Causal Modelling
DF	Discount Factor
DGC	Dynamic Granger Causality
DGM	Directed Graph Model
DLM	Dynamic Linear Model
DMN	Default Mode Network
Gen Synch	Generalised Synchronization
GES	Greedy Equivalence Search
ICA	Independent Component Analysis
IPA	Integer Programming Algorithm

LDS	Linear Dynamic System
LiNGAM	Linear Non-Gaussian Acyclic Model
LMDM	Linear Multiregression Dynamic Model
LPL	Log predictive likelihood
MAD	Mean Absolute Deviation
MAP	Maximum a Posteriori Probability
MDM	Multiregression Dynamic Model
MSE	Mean Squared Error
PCA	Principal Component Analysis
ROI	Region of Interest
SEM	Structural Equation Modelling
SHD	Structural Hamming Distance

Chapter 1

Introduction

In this thesis a class of Multiregression Dynamic Model (MDM) is applied to resting-state functional Magnetic Resonance Imaging (fMRI) data. Functional MRI consists of a dynamic acquisition, *i.e.* a series of images, which provides a time series at each volume element or voxel. These data are indirect measurements of blood flow, which in turn are related to neuronal activity. A traditional fMRI experiment consists of alternating periods of active and control experimental conditions, and the purpose is to compare brain activity between two different cognitive states (*e.g.* remembering a list of words versus just passively reading a list of words). In contrast, a "resting-state" experiment is conducted by having the subject remain in a state of quiet repose, and the analysis focuses on understanding the pattern of connectivity among different cerebral areas. The ultimate (and ambitious) goal is to understand how one neural system influences another (Poldrack *et al.*, 2011). Some studies assume that the connection strengths between different brain regions are constant. Dynamic models have been proposed for resting-state fMRI, but they usually estimate the temporal correlation between brain regions (rather than the influence that one region exerts to another) or their scores are not a closed form which complicates the process of learning network (see *e.g.* Chang and Glover, 2010; Allen *et al.*, 2012). However, clearly a more promising strategy would be to perform a search over a large class of models that is rich enough to capture the dynamic changes in the connectivity strengths that are known to exist in this application. The Multiregression Dynamic Model (MDM) can do just this (Queen and Smith, 1993; Queen and Albers, 2009), and in this thesis we demonstrate how it can be applied to resting fMRI.

1.1 Thesis Outline

This thesis begins with a discussion about the importance of investigating the brain when a person is in a state of rest (see Section 1.2). Then an introduction to fMRI is provided in Section 1.3. Chapter 2 clarifies the difference between the types of brain connectivity and provides a review of popular methods used to estimate the neural connections. Also, important definitions about graph theory are shown in Chapter 2.

Chapter 3 defines MDMs and gives a comparison between it and some other methods described previously in Chapter 2. To our knowledge, we present here the first application of Bayes factor MDM search. As with standard BNs, the Bayes factor of the MDM can be written in closed form, and thus the model space can be scored quickly. However unlike a static BN that has been applied to this domain, the MDM models *dynamic* links and so allows us to discriminate between models that would be Markov equivalent in their static versions. Furthermore, the directionality exhibited in the MDM graph can be associated with a causal directionality in a very natural way (Queen and Albers, 2009) which is also scientifically meaningful.

Even for the moderate number of variables needed in this application, the model space we need to search is extremely large; for example, a graph with just 6 nodes has over 87 million possible BNs, and for a 7 node graph there are over 58 billion (Steinsky, 2003). Instead of considering approximate search strategies, we exploit recent developments to perform a full search of the space, using the Integer Programming Algorithm (IPA; Cussens, 2011) for searching graphical model spaces. In Chapter 4, we then use synthetic data to demonstrate that the MDM-IPA is not only useful method for detecting the existence of brain connectivity, but also for estimating its direction. Another search method is presented in Chapter 3, called the MDM-DGM, which does not consider the acyclic constraints and searches the larger class of directed graphs. We apply both of these methods to fMRI datasets in Chapter 4.

Chapter 3 also presents new diagnostic methods customised to the needs of the MDM, analogous to those originally developed for static BNs, using the closed form of the one-step ahead predictive distribution (Cowell *et al.*, 1999). These diagnostic methods are essential because it is well known that Bayes factor model selection methods can break down whenever no member of the considered model class fits the data well. It is, therefore, important to check that selected models are consistent with the observed series.

We propose a strategy of using the MDM-IPA to search initially across a class of simple linear MDMs which are time homogeneous, linear and with no change points. We then check the best model using these new diagnostic methods. In practice, we have found the linear MDMs usually perform well for most nodes receiving inputs from other nodes. However, when diagnostics discover a discrepancy of fit, the MDM class is sufficiently expressive for it to be embellished to accommodate other anomalous features. For example, it is possible to include time-dependent error variances, change points, interaction terms in the regression and so on, to better reflect the underlying model and refine the analysis. Often, even after such embellishment, the model still stays within a conditionally conjugate class. Therefore, if our diagnostics identify serious deviation from the highest scoring simple MDM, we can adapt this model and its high scoring neighbours with features explaining the deviations. The model selection process using Bayes factors can then be reapplied to discover models that describe the process even better. In this way, we can iteratively augment the fitted model and its highest scoring competitors with embellishments until the search class accommodates the main features observed in the dynamic processes well. This is one advantage of adopting a fully Bayesian methodology to perform this analysis. Standard Bayesian diagnostics can be adapted to provide guidance in checking and where necessary to guide the modification of the model class. In Chapter 4, we demonstrate this process with real fMRI datasets.

FMRI experiments are usually conducted on more than one subject. Therefore, these studies need to take into account not only the interaction between areas of one single brain but also the differences among subjects. In Chapter 5, we present four approaches for estimating connectivity maps using a group of subjects. The first is the *virtual-typical-subject* (VTS) approach. Here a "typical subject" is identified as the average among the time series variables across subjects or simply concatenating all datasets. The second approach, called *common-structure* (CS), learns the same network for all individuals, but allows the connection strengths (regression parameters) between subjects to differ. The next approach, *individual-structure* (IS), learns a network for each subject and then the group network is defined as a combination of these individual networks. For the first time, to our knowledge, we develop the VTS, the CS and the IS approaches in the context of the MDM.

However, these group analysis methods cannot determine if the group of subjects is drawn from a single population, or from multiple populations with different connectivity

patterns. Thus, the next approach, *Group-structure* (GS), uses a cluster analysis to group homogeneous subjects according to their brain networks. Therefore, in Chapter 5, we also suggest a novel separation measure for comparing individuals based on the model selection measure, Bayes factor. Although the *group-structure* (GS) approach developed here is applied with the MDM, it can be used with any other graphical models. Comparing these approaches using synthetic and real fMRI data, we found that the GS approach provides results more scientifically consistent with the expected.

In Chapter 6, we develop the Marginal Estimation of Multiple Networks method, which estimates the individual and the group networks, considering the distance between them. We then present an extension of this method, called the Joint Estimation of Multiple Networks, developed initially by Oates *et al.* (2014), using a penalty function for dense graphs. We provide the first application of this method in real data, discussing some aspects in practice. Finally, Chapter 7 describes the directions for future work.

1.2 Resting-state

There is growing interest in the neuroscience literature about the brain at rest. Typical brain imaging studies have observed the behaviour of the brain when a person is doing a specific task, such as pressing buttons, speaking or even doing a mathematical calculation. In this way, these experiments have studied what Raichle (2010) called a "reflexive view of brain function" and have found important results that identify brain regions involved in various behaviours. Raichle (2010) argues that when researchers work only with task-evoked responses, they underestimate the function of the brain, and leave out the study of brain activity as "information processing for interpreting, responding to and predicting environmental demands". In simple terms, the brain continues to work even when the person is apparently not performing any activity, and in particular, the brain spends about 20% of the body's energy regardless of whether in a resting or task state.

How to study intrinsic activity

In a resting-state experiment, subjects remain at quiet repose with eyes closed, however, sometimes they may keep their eyes open with or without visual fixation. Initially, the resting-state data were measured with neuroimaging through positron emission topography

(PET) and more recently through functional magnetic resonance imaging (fMRI; see Section 1.3).

In 1997, Shulman *et al.* showed that the activity of some brain regions had decreased when the experiment changes from resting-state to goal-directed tasks, using PET (Figure 1.1(a)). This finding was confirmed later by Binder *et al.* (1999), Mazoyer *et al.* (2001) and Raichle *et al.* (2001). Initially, this finding was considered strange, because these brain areas did not form a known network, such as the motor or visual system. Hence, this new network was called the default mode network (DMN). Figure 1.1(a) illustrates two of the key brain areas in the DMN, the posterior cingulate cortex (yellow arrow) and the ventral medial prefrontal cortex (orange arrow). Using a resting-state fMRI experiment, Greicius *et al.* (2003) verified that these two regions (Figure 1.1(c)) have a similar activation pattern (Figure 1.1(b)).

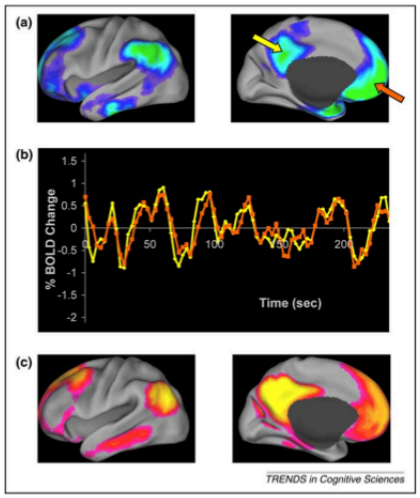


Figure 1.1: Illustration of the Default Mode Network (DMN). Panel (a) shows the areas that Shulman *et al.* (1997) found to be decrease during a task performance, the key areas being the posterior cingulate (yellow arrow) and ventral medial prefrontal cortex (orange arrow), eventually termed the DMN. Panel (c) shows the similar areas found when examining fMRI data collected during a resting-state (Greicius *et al.*, 2003), where Panel (b) shows fMRI time series for the two selected regions (yellow and orange lines for yellow and orange arrows in (a), respectively), showing the great similarity between these distant brain regions. (Figure from Raichle, 2010).

1.3 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a powerful tool that has been used to obtain resting-state data. In fact, the number of publications using fMRI data has grown exponentially. The fMRI's success, relative to other techniques like PET or electrophysiolog-

ical methods, is due its ability to noninvasively record brain function with good spatial and temporal resolution (Poldrack *et al.*, 2011). Also from a very practical standpoint, a fMRI experiment with *no task* is easier to carry out that one with a task.

What does fMRI measure?

FMRI data reflect the blood oxygenation level, which is indirectly related to the activation of brain neurons. When neurons increase their firing rate, they require more oxygen, which in turn results in an increase in blood flow in that region. Counterintuitively, the amount of oxygen delivered by the blood exceeds the increased demand for oxygen. Therefore, an increase in oxygen is indicative of activation of the neurons in that place. This change in oxygenation gives rise to the blood oxygenation level dependent signal (BOLD), the time series variable measured by fMRI.

The BOLD hemodynamic response (HR) is the temporal evolution of the fMRI signal induced by a change in neuronal activity (Poldrack *et al.*, 2011). A peculiar characteristic of the HR is the speed at which this measure responds to neural activity. Although changes in neuronal activity occur on the order of milliseconds, the HR only reaches its peak in about 5 seconds and then takes 15 to 20 seconds to return to baseline. Therefore, the BOLD does not react immediately to a stimulus because the blood flow changes slowly. For instance, Figure 1.2 shows the temporally delayed and blurred BOLD signal in response to a repeating on/off stimulus.

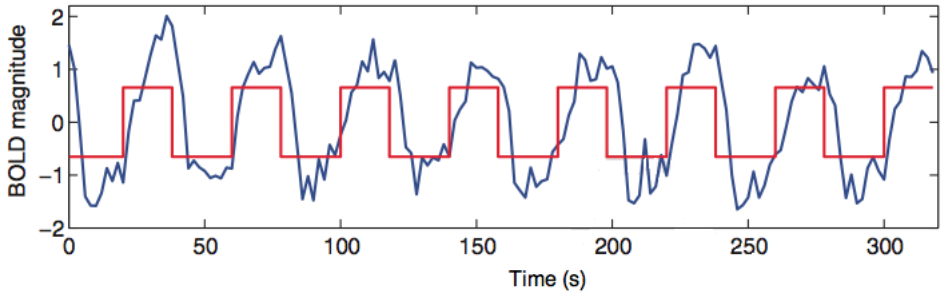


Figure 1.2: BOLD fMRI data (blue) for a block (on/off) experimental design (illustrated in red). Note the systematic temporal delay, with the BOLD data rising well after the start of each block, and falling well after the end of each block. Also, while there is some noise, the temporal BOLD signal has a smoother profile than the 'square wave' pattern of the experiment. (Figure from Poldrack *et al.*, 2011, Chapter 5).

Visualizing fMRI data

FMRI data takes the form of a time series of 3-dimensional images. The smallest point in the image is called a voxel, *i.e.*, it is like a pixel, but in 3 dimensions. FMRI data have arbitrary units and are typically visualised as a grayscale image. Each voxel has its spatial location in the brain that corresponds to three dimensions X, Y and Z, representing respectively, left-right, anterior-posterior and inferior-superior dimensions, as shown in Figure 1.3.

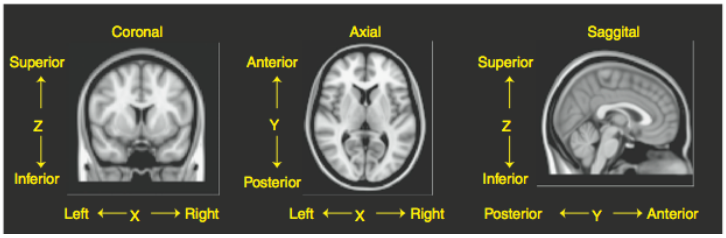


Figure 1.3: The three main axes used in the standard atlas space in neuroimaging. (Figure from Poldrack *et al.*, 2011, Chapter 2).

Analysis of resting-state fMRI data

Resting-state fMRI data is in some ways simpler to analyse than traditional task-design fMRI, principally because there is no task-related variability to model. However, there are critical preprocessing steps that are required for both resting and task fMRI.

The first preprocessing step is motion correction. While individuals are instructed to lie quietly in the scanner, they invariably move their head slightly which can result in dramatic signal changes. For example, for a voxel at the edge of the brain, even a small head movement could result in a 100% signal change. Rigid body image registration is used to align each volume to a reference (*e.g.* the first volume), thus removing this potentially large source of artifactual variation.

The next preprocessing step is inter-subject alignment. Different subjects' brains have different sizes and shapes, and hence registration is needed to align each subject to a common atlas space. The registration can be linear, *e.g.* an affine transformation (Jenkinson *et al.*, 2002), but more typically non-linear registration is used to 'warp' individual differences in brain anatomy to a common space (Ashburner & Friston, 2007). After inter-subject alignment, we can conduct analyses voxel-by-voxel, with reasonable assumption that a given voxel corresponds to the same brain region in each subject.

Analysis of resting-state fMRI data generally follows one of two approaches, region-based or voxel-wise (Hayasaka & Lurenti, 2010). A region of interest (ROI) approach reduces the dimensionality of the data using pre-defined anatomical regions. For example, a dataset with 500 time points and 100,000 voxels is reduced to 500 time points and 100 ROI's; the time series at each ROI is computed as the average of the intensities inside the ROI. Voxel-wise approaches often use multivariate exploratory methods, like Independent Components Analysis (ICA), to reduce dimensionality (Kiviniemi *et al.*, 2003). For example, with ICA the same 500 $\times$ 100,000 voxel dataset could also be reduced to 500 $\times$ 100 dataset, but the spatial patterns that define the 100 dimension are data-dependent. In Chapter 2, we discuss ICA and some other methods used to deal with the high dimensionality problem and also the individual variability.

Chapter 2

Brain Connectivity

A fundamental debate in neuroscience is whether specific brain areas are responsible for certain functions (functional segregation), or whether these activities are distributed over the entire brain (functional integration; Finger, 1994; Friston, 2011). When functional brain imaging methods like fMRI were first developed, the focus was on "mapping", *i.e.* a *functional segregation* approach. Experiments were conducted to identify the brain regions that changed systematically with the task. Functional segregation is supported by the large-scale organisation of the brain, for example, there is a "visual cortex", where visual information is processed, and a "motor cortex" involved in the control of movement. More recently, attention has focused on a *functional integration* approach, where a particular brain region may be responsible for several different functions depending on the pattern of interactions with other brain regions (Lenartowicz and McIntosh, 2005; Bressler and McIntosh, 2007; Sporns, 2011). As a result, presently there is much research on the intercommunication among brain regions, often just referred to as the study of *connectivity* (Friston, 2005; McIntosh, 2000).

The study of brain networks focuses on how the interaction between different brain regions guides thought, behaviour, consciousness, learning and so on. Therefore through connectivity study it is possible to identify sets of brain regions where a particular function is localised and to suggest which neural systems are involved in this process. Moreover, as some diseases or damage may cause changes in the connectivity pattern, it helps to predict the consequences of the changes or to find the best way to recovery from them.

In this chapter we present the three types of connectivity: structural, functional and effective, and some methods used to estimate them. We pay particular attention to Bayesian

Network (BN) models and the graphical theory underpinning this approach.

2.1 Structural, Functional and Effective Connectivity

The study of connectivity can be focused on two aspects: the segregation of network into regions (local connectivity) and the network-wide integration (global connectivity; Sporns, 2011). Some brain areas are strongly interconnected with each other, while others are less interconnected. Brain networks may be formed on the basis of anatomical links or from statistical or causal relationships among brain areas.

Structural connectivity concerns the anatomical links between brain elements, for example, neuron systems or interregional pathways. These connections are regarded as static over the time frame of data acquisition, though they can change over the lifespan. The detailed knowledge of structural connectivity, albeit important, is insufficient to infer on the complex changes in brain function (Sporns, 2011). As an analogy, structural connectivity is like a traditional (paper) map; the map can tell you the paths and the sizes of the roads, but not how the traffic is flowing at any particular instant.

Functional integration concerns how different parts of the brain work together to yield behaviour and cognition. Two broad distinctions are made in studies of functional integration, between *functional connectivity* and *effective connectivity*. The former is defined as correlation or statistical dependence among the measurements of neuronal activity of different areas, while the latter corresponds to a direct causal influence (Friston, 2011). Of course, a significant correlation between the two regions does not imply that one region directly influences the other one. For instance, the following three situations may lead to the existence of functional connectivity between the two regions (Poldrack *et al.*, 2011). First, a region may indeed directly affect another region, and then there is an effective connectivity between them, *e.g.* regions *1* and *3* in Figure 2.1. Another possible scenario is when there is a significant correlation between the regions *1* and *4*, because an indirect influence of *3*, which is directly influenced by *1* and also influences *4*. Finally, a third situation is when a single region *3* influences two other regions, *4* and *5*. Thus, an activation in this region *3* leads to a response in both regions, making this region *3* responsible for the correlation between regions *4* and *5*. Studies have been developed to define and detect a direct influence between variables, especially in the area of machine learning (see e.g. Spirtes

et al., 2000 and Pearl, 2000). Friston (2011) asserted that effective connectivity can be seen as a temporal dependence between brain areas and therefore it may be defined as dynamic (activity-dependent). In Section 2.3, we review some methods used to estimate functional and effective connectivity.

Some studies have shown a significant positive correlation between the pairwise structural and functional connections (Honey *et al.*, 2009). In addition, while structural connections can inform functional connections, functional connections are not good predictors of structural coupling, because there may be a functional connection between brain regions that are only indirectly anatomically linked (Sporns, 2011). Therefore, it is not possible to completely understand the brain network using only one mode of connectivity. For instance, a particular task may evoke effective connectivity amongst a set of brain regions and then the structural connectivity may be used to find more complete interpretation based on biophysical mechanisms (Sporns, 2011).

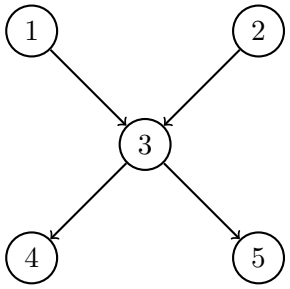


Figure 2.1: A graphical structure considering 4 nodes.

2.2 Some Definitions of Graphical Modelling

The brain connectivity is usually studied through a graphical model in which the causal relationships are expressed in terms of conditional independence among random variables (*nodes*), considering a graphical structure. A graph consists of nodes and *edges* in which the latter represents the connection between pairs of the former. In connectivity studies, nodes can be considered as voxels. Or they can be defined as segregated brain regions, found through a cluster analysis used to group homogeneous voxels (Sporns, 2011); these regions are usually referred to as Regions of Interest (ROIs). After that an integration study can be used to find edges, *e.g.* Bayesian Network, see below. When edges are not ordered, *i.e.*, the edge from node i to node j is identical to the edge from node j to node i, the graph is formed

by *undirected* edges. When one node influences other, the edge indicates the direction of the effect. In the example shown in Figure 2.1, the edges are *directed* and the network is called a *digraph*. When there is a directed edge from one node to another, the former is called a *parent* while the latter is a *child*. For instance, node *1* and node *2* are the parents of node *3* and so it is a child in Figure 2.1. Moreover *neighbours* are nodes connected by an edge. The family of a node is called its *Markov blanket*, which consists of its parents, its children and the parents of its children. A node that does not have a parent is called *root* node and it represents original causes, *e.g.* nodes *1* and *2*. A *leaf* node does not have children, and it represents final effects, *e.g.* nodes *4* and *5*. If a node has both parent and children, it is called an *intermediate* node, *e.g.* node *3*.

In terms of notation, a graph or network G can be defined by its set of nodes ($\mathcal{N}$) and edges (ϵ), G = G($\mathcal{N}$, ϵ). The graph can also be defined by its adjacency matrix A, G = G($\mathcal{N}$, A). An *adjacency matrix* or connection matrix is a square matrix with binary elements representing the presence or absence of edges. This matrix is symmetric for undirected graphs and asymmetric for directed graphs. The total number of edges is $E = \sum_{i>j} A_{ij}$ in an undirected graph and $E = \sum_{ij} A_{ij}$ in a directed graph, where A_{ij} is the $(i, j)^{th}$ element of matrix A. The *degree* of a node is the number of edges connected to this particular node in an undirected graph. In directed graphs, the number of edges that leave from and arrive at a node are respectively called the *outdegree* and *indegree*. Nodes with high outdegree have the control of network system whilst nodes with high indegree are more affected by others. The adjacency matrix does not have to be binary, and it can have elements that represent the weight of the edges. In this case the sum of all edge weights provides a similar measure of degree (Sporns, 2011).

Another type of graph is the *directed acyclic graph* (DAG), which means that no path starts and ends at the same node, *i.e.* it exhibits no *cycles*. A *path* is a ordered sequence from one node to another passed by edges and intermediate nodes. One node is an *ancestor* of another if the former belongs to any path between a root and the latter node. In contrast, a node is a *descendant* of another if the former belongs to any path between the latter and a leaf node. A path is said to be *blocked* by a set of nodes, say $\mathbf{W}$, if the path contains:

- a *chain*, $x \rightarrow y \rightarrow z$, where $y \in \mathbf{W}$, or
- a *fork*, $x \leftarrow y \rightarrow z$, where $y \in \mathbf{W}$, or

- a *collider*, $x \rightarrow y \leftarrow z$, where $y \notin \mathbf{W}$ and no descendant of y is in $\mathbf{W}$.

Two disjoint sets of nodes, say $\mathbf{U}$ and $\mathbf{V}$, are said to be *d-separated* by $\mathbf{W}$ if any element of $\mathbf{W}$ blocks every path between $\mathbf{U}$ and $\mathbf{V}$.

2.3 Some Methods for Discovering Connectivity

In this section, we describe some techniques that have been used to estimate connectivity. After describing the MDM in Chapter 3, we discuss theoretically the similarities and differences among some of these methods and the MDM in Section 3.4.

Seed voxel correlation

Although it is interesting to study the correlation between all parts of the brain, sometimes this is not feasible for fMRI data. Generally, there is data on more than 10,000 voxels, which means millions of possible pairwise correlations. To address this, seed-based studies begin by defining "seed" regions, then compute the average time series for the seed region, and finally correlate the seed time series with time series at each voxel in the brain. In some variants, a set of seed regions is used, and only correlations amongst these regions are considered (Biswal *et al.*, 1995; Cordes *et al.*, 2000; Fox *et al.*, 2005). However, Varoquaux and his collegues (2010) pointed out that despite practicality of this approach, the resulting analyses obviously depend on the seed regions chosen. For this reason, clustering techniques have also been developed to study the connectivity amongst all brain regions, without the need to choose such seed regions. The most popular way of doing this in the field of fMRI is to perform an Independent Component Analysis.

Independent Component Analysis - ICA

This technique models a multivariate signal as a linear function of independent source signals. If 3-dimensional space is "unwrapped", then fMRI data can be represented as a time $\times$ space matrix A. The ICA model takes the form $A = MS$, where M is a mixing matrix (time $\times$ components) and S is unknown independent sources (components $\times$ space). ICA estimates the matrices M and S by maximising the pairwise (statistical) independence of each row of S. Usually a Principal Component Analysis (PCA) is performed before an ICA

in order to reduce the dimension of the data (see *e.g.* Beckmann and Smith, 2004, who also introduced a noise model into a PCA). Typical algorithms minimize the Mutual Information or maximize the non-Gaussianity measurement (McKeown *et al.*, 1998; Hyvärinen, 2000; Kiviniemi *et al.*, 2003). As ICA is a data-driven exploratory method, the independent components given by ICA vary from one dataset to other. Nevertheless with this caveat, ICA can be a powerful tool, for example, for the classification of groups and mapping brain patterns according to a diagnostic class (Varoquaux *et al.*, 2010). Indeed, it is widely used with resting-state data, where there are few prior hypotheses to be tested. However, ICA cannot be used to compare models and test hypotheses about directed causal influences between brain regions (Friston, 2011). In this thesis, we are using the ICA to define the ROIs as a preprocessing step, but some approaches, for example the Linear Non-Gaussian Acyclic Model (see below), incorporate the ICA in their methods for estimating connectivity.

Full correlation and partial correlation

The simplest way to study the relation among brain areas is through the covariance or correlation matrix of their corresponding time series. Each time series may correspond to a single voxel, an average over an ROI, or an ICA temporal component (a column of M). Full correlation (so called to distinguish from partial correlation) does not distinguish direct from indirect connections. In contrast, partial correlation can inform the direct relation between two variables, after allowing for the effect of other variables. Thus the partial correlation between two variables X and Y given a set of variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$ is calculated as the correlation between the residuals $r_{X.Z}$ and $r_{Y.Z}$, where these residuals are found through linear regressions where X and Y are respectively dependent variables and $\mathbf{Z}$ is the controlled variable (Baba *et al.*, 2004; Marrelec *et al.*, 2006). Note that full and partial correlation estimate functional connectivity with symmetric relation among brain areas. Therefore, it is not possible to estimate the directional relationship between regions.

Patel's conditional dependence measures

A simplified approach for estimating connectivity was proposed by Patel *et al.* (2006) based on a comparison between conditional and marginal probability of elevated activity. For a pair of brain regions, this method starts by binarising the two time series with an arbitrary

threshold, producing a sequence of "elevated activity" measurements.

Patel *et al.* (2006) then calculated:

$$\begin{aligned} Z_{ij1} &= \sum_{s=1}^{S}\sum_{t=1}^{T} I(Y^*_{st}(i)=1, Y^*_{st}(j)=1), \\ Z_{ij2} &= \sum_{s=1}^{S}\sum_{t=1}^{T} I(Y^*_{st}(i)=1, Y^*_{st}(j)=0), \\ Z_{ij3} &= \sum_{s=1}^{S}\sum_{t=1}^{T} I(Y^*_{st}(i)=0, Y^*_{st}(j)=1), \\ Z_{ij4} &= \sum_{s=1}^{S}\sum_{t=1}^{T} I(Y^*_{st}(i)=0, Y^*_{st}(j)=0), \end{aligned}$$

where I is a indicator variable and $Y^*_{st}(i)$ is a dichotomized variable that represents whether region i is active at time t for subject s. The discrete variable $\mathbf{Z}_{ij} = (Z_{ij1}, Z_{ij2}, Z_{ij3}, Z_{ij4})$ follows a Multinomial distribution with parameter $\boldsymbol{\alpha}_{ij} = (\alpha_{ij1}, \alpha_{ij2}, \alpha_{ij3}, \alpha_{ij4})$, where

$$\begin{aligned} \alpha_{ij1} &= p(Y^*_{st}(i)=1, Y^*_{st}(j)=1), \\ \alpha_{ij2} &= p(Y^*_{st}(i)=1, Y^*_{st}(j)=0), \\ \alpha_{ij3} &= p(Y^*_{st}(i)=0, Y^*_{st}(j)=1), \\ \alpha_{ij4} &= p(Y^*_{st}(i)=0, Y^*_{st}(j)=0), \end{aligned}$$

and $\boldsymbol{\alpha}_{ij}$ assumes a Dirichlet prior distribution. The variable Z_{ij1} can be interpreted as the number of times that regions i and j showed an elevated activity at the same time.

Then, a measure κ_{ij} is evaluated as

$$\kappa_{ij} = \frac{\alpha_{ij1} - E}{W(\min(\alpha_{ij1}+\alpha_{ij2}, \alpha_{ij1}+\alpha_{ij3}) - E) + (1-W)(E - \max(0, 2\alpha_{ij1}+\alpha_{ij2}+\alpha_{ij3}-1))},$$

where $E = (\alpha_{ij1}+\alpha_{ij2})(\alpha_{ij1}+\alpha_{ij3})$ and $W = \frac{\alpha_{ij1}-E}{2(\min(\alpha_{ij1}+\alpha_{ij2}, \alpha_{ij1}+\alpha_{ij3})-E)} + 0.5$, if $\alpha_{ij1} \geq E$, or $W = 0.5 - \frac{\alpha_{ij1}-E}{2(E-\max(0, 2\alpha_{ij1}+\alpha_{ij2}+\alpha_{ij3}-1))}$, otherwise.

This is a measure of association that compares the estimated value of the joint probability $p(Y^*_{st}(i), Y^*_{st}(j))$ with its expected value under the independence assumption, *i.e.* $p(Y^*_{st}(i), Y^*_{st}(j)) = p(Y^*_{st}(i))p(Y^*_{st}(j))$. This measure is found for each pair of brain areas, and varies between -1 and 1. When $\kappa_{ij} = 0$, the joint distribution and the product of the marginal probabilities are the same and, in this sense, it can be concluded that regions i and

j are not connected.

Finally, when two particular brain regions are connected (*i.e.* $\kappa_{ij} \neq 0$), measure τ_{ij} is calculated based on the ratio of the marginal probabilities of each region. When $\tau_{ij} > 0$, the region i is ascendant to the region j whilst the negative value of this measure means that the region j is ascendant to the former region. By definition, the node j is called ascendant to node i if the marginal activation probability of the former node is larger than that of node i.

In contrast to the other measures of association for 2×2 tables, such as Cohen's Kappa (Cohen, 1960) and Mutual Information (Cover and Thomas, 1991), Patel *et al.* (2006) treats the marginal activation probabilities as fixed, and then it is possible to estimate the ascendancy. Therefore, using the binary variables, it is possible to estimate the directed link between a pair of brain regions in a simple way, and so, the connectivity here is estimated based on the joint probabilities of elevated activity. However, of course, there is a loss of information when continuous variables are transformed into binary, and possibly because of this, Smith, S.M. *et al.* (2011) found that this approach had a poor performance in detecting the presence of a network connection, using the synthetic data. Moreover, this approach provides only static estimates for connectivity, and so, it is not possible to verify whether connectivity changes over time.

Generalised synchronization (Gen Synch)

Synchronization phenomena are also studied to investigate the communication between different brain areas (Quian Quiroga *et al.*, 2002; Pereda *et al.*, 2005; Dauwels *et al.*, 2010). In the approach proposed by Arnhold *et al.* (1999), each time series is embedded in a high-dimensional state-space, specifically, by creating multivariate time series from univariate ones. Specifically, if $\mathbf{Y}(i) = (Y_1(i), \ldots, Y_T(i))$ is the time series measured for region i, then $\mathbf{Y}_t^*(i) = (Y_t(i), Y_{t-1}(i), \ldots, Y_{t-(m-1)\tau}(i))$, where m is the embedding dimension (*e.g.* m=10) and τ is the time lag. Define $r_{tk}(i)$ as the time indices of the k nearest neighbours of $\mathbf{Y}_t^*(i)$, for $k = 1, \ldots, K$ so that

$$
\begin{aligned}
||\mathbf{Y}_t^*(i) - \mathbf{Y}_{r_{t1}(i)}^*(i)|| &= \min_q ||\mathbf{Y}_t^*(i) - \mathbf{Y}_q^*(i)||, \\
||\mathbf{Y}_t^*(i) - \mathbf{Y}_{r_{t2}(i)}^*(i)|| &= \min_{q \neq r_{t1}(i)} ||\mathbf{Y}_t^*(i) - \mathbf{Y}_q^*(i)||,
\end{aligned}
$$

and so on, where $||\mathbf{A} - \mathbf{A}'||$ is the Euclidean distance.

Therefore, the mean squared Euclidean distance between $\mathbf{Y}_t^*(i)$ and its the k nearest neighbours is:

$$R_k(\mathbf{Y}_t^*(i)) = \frac{1}{K}\sum_{k=1}^{K} ||\mathbf{Y}_t^*(i) - \mathbf{Y}_{r_{tk}(i)}^*(i)||^2,$$

whilst the conditional mean squared Euclidean distance of $\mathbf{Y}_t^*(i)$, conditioned on the k nearest neighbours of other region, say $\mathbf{Y}_t^*(j)$, is:

$$R_k(\mathbf{Y}_t^*(i)|\mathbf{Y}_t^*(j)) = \frac{1}{K}\sum_{k=1}^{K} ||\mathbf{Y}_t^*(i) - \mathbf{Y}_{r_{tk}(j)}^*(i)||^2.$$

According to Arnhold *et al.* (1999), $R_k(\mathbf{Y}_t^*(i)|\mathbf{Y}_t^*(j)) \gg R_k(\mathbf{Y}_t^*(i))$, when the time series variables of regions i and j are completely independent. Therefore, an interdependence measure was defined as:

$$S_k(i,j) = \frac{1}{T}\sum_{t=1}^{T} \frac{R_k(\mathbf{Y}_t^*(i))}{R_k(\mathbf{Y}_t^*(i)|\mathbf{Y}_t^*(j))},$$

where $0 < S_k(i,j) \leq 1$, and the small values of $S_k(i,j)$ indicate independence. However, Arnhold *et al.* (1999) also suggested other measure based on a geometrical average as follows,

$$\begin{aligned} H_k(i,j) &= \frac{1}{T}\sum_{t=1}^{T} \log \frac{R(\mathbf{Y}_t^*(i))}{R_k(\mathbf{Y}_t^*(i)|\mathbf{Y}_t^*(j))}, \text{ where} \\ R(\mathbf{Y}_t^*(i)) &= \frac{1}{T-1}\sum_{l\neq t} ||\mathbf{Y}_t^*(i) - \mathbf{Y}_l^*(i)||^2, \end{aligned}$$

and so Quian Quiroga *et al.* (2000) showed that this measure $H_k(i,j)$ is more robust against noise than $S_k(i,j)$ using a couple chaotic systems study. However, $H_k(i,j)$ is not normalised, assuming positive or negative values. Its interpretation is the same as for $S_k(i,j)$, *i.e.* if the time series variables are completely independent, then $H_k(i,j) = 0$.

Quian Quiroga *et al.* (2002) suggested an alternative measure, using also a different way of averaging, as

$$N_k(i,j) = \frac{1}{T}\sum_{t=1}^{T} \frac{R(\mathbf{Y}_t^*(i)) - R_k(\mathbf{Y}_t^*(i)|\mathbf{Y}_t^*(j))}{R(\mathbf{Y}_t^*(i))}.$$

In theory, Quian Quiroga *et al.* (2002) showed that $N_k(i,j)$ has better characteristics than the previous measures, *i.e.*, it is normalised (although it may provide slightly negative values)

and its interpretation is the same as for $S_k(i,j)$, but, similar to $H_k(i,j)$, in principle $N_k(i,j)$ is more robust than $S_k(i,j)$. However, in practice, Quian Quiroga *et al.* (2002) found similar results when these measures were applied to real datasets, and also, Smith, S.M. *et al.* (2011) reported that $H_k(i,j)$ and $N_k(i,j)$ provide similar results using synthetic data. Similar to Patel's measures, Gen Synch also provides static estimates for connectivity, and also did not show a good performance in detecting the presence of a network connection, in the study of Smith, S.M. *et al.* (2011).

The Time Varying Undirected Graph

The time varying undirected graph (TVUG) supposes that n-dimensional time series at time t, $\mathbf{Y}_t$, follows a multivariate Gaussian distribution with mean zero and the covariance matrix $\Sigma(t)$. These variables, $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$, are assumed to be independent over time but not identically distributed (Zhou *et al.*, 2010). Therefore, the graphical structure changes over time based on $\Sigma(t)$ so that the edges correspond to the elements different from zero in the inverse covariance matrix. In addition, the TVUG assumes that $\mathbf{Y}_t$ changes smoothly, and so it considers the model:

$$\mathbf{Y}_t^* = \mathbf{Y}_{t-1}^* + \mathbf{Y}_t,$$

where $\mathbf{Y}_0^* \sim \mathcal{N}(\mathbf{0}, \Sigma(0))$ and $\mathbf{Y}_t \sim \mathcal{N}(\mathbf{0}, \Sigma(t))$. Zhou *et al.* (2010) proposed an estimate of the covariance matrix at time t ($\Sigma(t)$) based on the ℓ_1-penalized maximum likelihood estimator (Banerjee *et al.*, 2008; Friedman *et al.*, 2008), thus

$$\hat{\Sigma}(t) = \arg\min_{\Sigma} \operatorname{tr}(\Sigma^{-1}\hat{S}_c(t)) + \log|\Sigma| + \lambda_c|\Sigma^{-1}|_1,$$

$$\text{where} \hat{S}_c(t) = \frac{\sum_{s=1}^{T} w_{st}\mathbf{Y}_s\mathbf{Y}_s'}{\sum_{s=1}^{T} w_{st}}$$

is a kernel estimator of the covariance, with weights $w_{st} = K(\frac{|s-t|}{h_T})$ given by a symmetric nonnegative function kernel over time (h_T can be defined as $T^{-1/3}$), Σ is a symmetric and positive definite matrix, $\operatorname{tr}(\mathbf{A})$ is the trace of a matrix $\mathbf{A}$, $|\mathbf{A}|$ denotes the determinant of a matrix $\mathbf{A}$, $|\mathbf{A}|_1$ is the ℓ_1 norm of a matrix $\mathbf{A}$, and λ_c is a non-negative regularisation parameter, which may be defined in a cross-validation study (for example, considering the value that minimises the likelihood loss, as shown by Banerjee *et al.*, 2008).

In contrast to most of methods used to estimate connectivity, the TVUG allows the connectivity changes over time. However, it uses the covariance matrix, and methods that are based on the second-order statistics are not usually able to estimate precisely the full causal structure (see *e.g.* Shimizu *et al.*, 2006).

Linear Non-Gaussian Acyclic Model - LiNGAM

A Linear Non-Gaussian Acyclic Model (LiNGAM) is used to estimate effective connectivity, based on the following assumptions: (1) data are generated through a linear process consistent with an acyclic graphical structure; (2) there are no unobserved confounders; (3) noise variables are mutually independent and have non-Gaussian distributions with non-zero variances (Shimizu *et al.*, 2006). Thus suppose $\mathbf{Y}$ is the observed (regions $\times$ time) data matrix. Then LiNGAM consists of the model:

$$\mathbf{Y} = \mathbf{B}\mathbf{Y} + \mathbf{e};$$

where $\mathbf{B}$ is a lower triangular matrix with all zeros on the diagonal and $\mathbf{e}$ is a residual matrix. Solving for $\mathbf{Y}$, we obtain the equation $\mathbf{Y} = \mathbf{A}\mathbf{e}$, where $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$ and $\mathbf{I}$ is the identity matrix. Because the components of $\mathbf{e}$ must be mutually independent and non-Gaussian, $\mathbf{A}$ can be identified through ICA. In general, the mixing matrix of ICA cannot be determinate under the assumption of Gaussianity, because many different mixing matrices yield the same covariance matrix, and so the same Gaussian joint density. Therefore, the assumption of non-Gaussianity enables the direction of relationships to be identified so that the effective connectivity can be estimated (Shimizu *et al.*, 2006). In practice, S.M. *et al.* (2011) showed that this method had a poor performance in detecting the presence of connectivity, and also in distinguishing its directionality. In addition, note that the LiNGAM only provides static estimates.

Structural Equation Modelling - SEM

The SEM is a generalization of a multiple regression model but with a flexibility over some of the assumptions. A structural equation may be written as:

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\theta}_t + \mathbf{e}_t,$$

where $\mathbf{Y}_t$ is an n-dimensional vector of variables observed at time t. The $n \times p$ matrix $\mathbf{X}_t$ contains p predictor variables which can be some of the response variables, if they are considered potentially dependent on one another, and also exogenous variables, *i.e.* its values are considered conditionally fixed in the model. The p-dimensional vector of parameters is $\boldsymbol{\theta}_t$, and $\mathbf{e}_t$ is an n-dimensional error vector which are not necessarily assumed to be independent of each other.

Thus, the observed correlation matrix of the variables is compared with the estimated correlation matrix of the best fitting model (for more details see Bollen and Long, 1993; Hoyle, 1995; Kline, 2010). Friston (2011) asserted two criticisms of this method. The first is that, because SEM models are most usually used to analyse a system in equilibrium, the SEM may not be well-suited for time series data. Second it is difficult to estimate the cyclic connections as even with the usual Gaussian assumption, the likelihood of the parameters of such models is very complicated.

Dynamic Granger Causality - DGC

As mentioned above, effective connectivity is sometimes interpreted as a measure of the direct causal influence that one neural activity can exert on another. Thus, some authors have discussed the definition of causality and how to measure it. In 1956, Wiener suggested a basic idea of causality saying that time series i has a causal influence on time series j when the prediction of j becomes better given the knowledge of the past of i. Almost ten years later, Granger (1969) used a linear regression model to implement Wiener's idea. More formally within this paradigm, the causal influence is judged to exist when the inclusion of past measurements from one time series reduces the variance of the autoregressive prediction error of another series (for more details, see Ding *et al.*, 2006 who showed the mathematical formalism in both time and spectral domain).

The multivariate autoregressive (MAR) model uses the idea of Granger causality to estimate the interaction amongst brain areas (see *e.g.* Yamashita *et al.*, 2005):

$$\mathbf{Y}_t = \sum_{l=1}^{L} \mathbf{A}_l \mathbf{Y}_{t-l} + \mathbf{v}_t,$$

where L is the MAR order, $\mathbf{A}_l$ is the $n \times n$ matrix that represents the connectivity between

the past with lag l and current observation variables, $\mathbf{Y}_t$, n is the number of variables, and $\mathbf{v}_t$ is the n-dimensional white Gaussian error with zero-mean and variance V.

Havlicek *et al.* (2010) developed a dynamic version of MAR. They transformed the matrix $\mathbf{A}_l$ into a time-varying connectivity $\mathbf{A}_l(t)$ and included a system equation as follows.

$$\begin{aligned} \mathbf{Y}_t &= \sum_{l=1}^{L} \mathbf{A}_l(t)\mathbf{Y}_{t-l} + \mathbf{v}_t, \\ \mathbf{a}_t &= \mathbf{a}_{t-1} + \mathbf{w}_t, \qquad\qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t), \end{aligned}$$

where $\mathbf{a}_t = \text{vec}([\mathbf{A}_1(t), \ldots, \mathbf{A}_L(t)]')$ and $\mathbf{w}_t$ is innovation at time t with the state variance $\mathbf{W}_t$. Havlicek *et al.* (2010) then extended their model to the frequency domain using the generalized partial directed coherence (GPDC). They could then test the significance of connectivity through the multivariate bootstrap-based approaches. These dynamic autoregressive models allow cyclic dependencies, but are very sensitive to the particular sampling rate. Also, model selection over the full model space is very complex because the dimension parameter space grows exponentially with maximal AR lag.

Classes like this one that directly model Granger causality have received severe criticism when applied to the fMRI datasets (Chang *et al.*, 2008; David *et al.*, 2008; Valdés-Sosa *et al.*, 2011; Smith *et al.*, 2012). For instance, the match between the sampling interval and the time constant in the neurodynamics is often poor, because the temporal delay blood-based response can vary considerable across brain regions. In fact, Smith, S.M. *et al.* (2011) discovered that lag-based approaches like these do not perform well at identifying connections for fMRI data, albeit only under the assumption of static connectivity strength.

The Linear and Bilinear Dynamic System

Other much more sophisticated classes of state space models have also recently been developed to model effective connectivity. These include the Linear Dynamic System (LDS; Smith *et al.*, 2010; Smith, J.F. *et al.*, 2011) and the Bilinear Dynamic System (BDS; Penny *et al.*, 2005; Ryali *et al.*, 2011). Smith, J.F. *et al.* (2011) defined the LDS as

$$\begin{aligned} \mathbf{Y}_t &= \boldsymbol{\beta}\boldsymbol{\Phi}\mathbf{s}^{t/\{t-L\}} + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}); \\ \mathbf{s}_t &= \mathbf{A}_{u_t}\mathbf{s}_{t-1} + \mathbf{D}_{u_t}\mathbf{h}_t + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_{u_t}); \end{aligned}$$

where the observed fMRI signal ($\mathbf{Y}_t$) is written as a function of the parameter $\boldsymbol{\beta}$ that represents the weight of a known convolution matrix $\boldsymbol{\Phi}$, and the past at lag L of the latent variables, *i.e.* the quasi-neural level variables, $\mathbf{s}^{t/\{t-L\}} = (\mathbf{s}'_{t-L}, \ldots, \mathbf{s}'_t)'$, $\mathbf{s}_t = (s_t(1), \ldots, s_t(n))'$; $\mathbf{v}_t$ is additive white Gaussian error. The matrix $\mathbf{A}_{u_t}$ represents the relationships among the latent variables, and is, therefore, responsible for estimating the effective connectivities whilst the matrix $\mathbf{D}_{u_t}$ is the set of regression coefficients of driving inputs (h_t) on the latent variables; u_t indexes the different connectivity states over the duration of the experiment. In a BDS, $\mathbf{A}_{u_t} = \mathbf{A} + \mathbf{B}\boldsymbol{\Lambda}_t$, where $\mathbf{A}$ indicates the interactions among latent variables without considering the influence of the experimental condition whilst the $\mathbf{B}$ represents the connections in the presence of modulatory inputs ($\boldsymbol{\Lambda}_t$). To estimate connectivity, these methods need to use approximate inferential methods, which complicates the search network process over a large model space. Moreover, the LDS and the BDS consider connectivity as static or estimate only the different strengths of connectivity when modelling a different experimental situation.

The Dynamic Causal Modelling - DCM

Another popular approach in the neuroscience literature estimates effective connectivity using Dynamic Causal Modelling (DCM; Friston *et al.*, 2003; Stephan *et al.*, 2008). DCM was developed for BOLD fMRI summarised in ROIs. The method poses a directed graphical model for unobservable neuronal populations, one per ROI, and detailed biophysical model to connect the neuronal activity to the measurable fMRI data. At each ROI there the key state variable represents neuronal activity ($\mathbf{s}$), and there are 4 other state variables for the haemodynamic response. There are also 5 (static) haemodynamic parameters estimated with the help of highly informative priors. See Friston *et al.* (2000) for full details.

The dynamics of neuronal states are assumed to evolve according to some equations of motion, such as

$$\dot{\mathbf{s}} \approx (\mathbf{A} + \sum_j u_j \mathbf{B}_j)\mathbf{s} + \mathbf{D}\mathbf{u}.$$

The effective connectivity matrix $\mathbf{A}$ represents the relationships among the brain regions without the influence of inputs $\mathbf{u} = \{u_1, \ldots, u_J\}$. In contrast, $\mathbf{B}_j$ represents the change in coupling due the *jth* input. Finally, the parameter $\mathbf{D}$ represents the direct influence of the experiment on neuronal activity.

The deterministic DCM assumes that the latent variables are completely determined by the model, *i.e.* the state variance is considered to be zero. As this version of DCM does not consider the influence of random fluctuation in neuronal activity, it cannot be used for resting-state connectivity (Penny *et al.*, 2005; Smith, J.F. *et al.*, 2011). More recently a stochastic DCM has been developed that addresses this problem (*e.g.* Daunizeau *et al.*, 2009 and Li *et al.*, 2011). In this model, the neuronal states are written as

$$\begin{aligned} \dot{\mathbf{s}} &\approx (\mathbf{A} + \sum_j u_j \mathbf{B}_j)\mathbf{s} + \mathbf{D}\mathbf{v} + \boldsymbol{\omega}^{(s)}, \\ \mathbf{v} &= \mathbf{u} + \boldsymbol{\omega}^{(v)}, \end{aligned}$$

where both random state fluctuations $\boldsymbol{\omega}^{(s)}$ and $\boldsymbol{\omega}^{(v)}$ follow a Gaussian distribution.

Both versions of the DCM depend on a nonlinear biophysical "Balloon model", making the inference process quite complex and infeasible for more than just a few nodes (Stephan *et al.*, 2010; Poldrack *et al.*, 2011). As such, it is cannot be used for most of the applications in this thesis. Furthermore, several authors have criticised the use of the Balloon model as speculative (Roebroeck *et al.*, 2011; Ryali *et al.*, 2011), which also make the use of these models less attractive.

Smith, S.M. *et al.* (2011) have recently compared most of these methods cited above — but not the MDM — using synthetic fMRI data. They found that approaches using BNs (defined in the next section) were found to provide the best results in detecting the presence of a network connection, whilst Patel's measures and Gen Synch (see above) appeared to be the best methods in distinguishing the directionality of the relation between the brain regions. We used the same synthetic data from Smith, S.M. *et al.* (2011) to compare some of these approaches with the MDM in Chapter 4.

2.4 Bayesian Network

2.4.1 What is a Bayesian Network?

A Bayesian Network (BN) models a stochastic process through a set of random variables whose conditional distributions are related via a graph structure (Smith and Croft, 2003; Dehmer, 2011). This focus of BNs has made them a standard data analysis tool in diverse

scientific fields, including medical diagnosis (Heckerman, 1990), learning maps (Dean, 1990), language interpretation (Goldman, 1990), protein networks (Kim *et al.*, 2006) and communication networks (Gibbens, 2000). Moreover, in recent years there has been growing interest in online "networking communities", *e.g.* Facebook and LinkedIn (Cohen, 2007).

One of the main strengths of Bayesian Networks is that their structure can accommodate any relevant information regarding when and how the modeled variables relate to each other, including the representation of candidate causal relationships (Sucar, 2006). A second strength of the Bayesian Network paradigm, particularly relevant to realistically complex application frameworks, is that depending on the purpose of the model and data availability a modeler might choose to use static or time-dependent network structures. The former class of models considers a single time slice to study the set of links, whereas the latter, commonly known as the dynamic bayesian network (DBN, see below) studying the changes in the relationship between the variables over time (Goldenberg, 2009).

2.4.2 The Joint Probability Distribution of the BN

Another definition of Bayesian Network models is that they decompose the joint distribution of a set of observables into a set of conditional distributions. BNs embody the assumption of the Markov property (see below), and only considers direct dependencies that are explicitly shown via edges (Korb and Nicholson, 2004). For instance, in the graph in Figure 2.1, node *1* only influences node *4* through node *3*; in other words, there is no hidden link from node *1* to node *4*. A *sparse* BN, which has few parents for each node, has a computationally tractable joint probability distribution.

In order to better understand the relation between conditional independence and BNs, Korb and Nicholson (2004) study some different graph structures. For instance, consider the *causal chain* formed by nodes *1*, *3* and *4* in Figure 2.1. In this case, considering the respectively random variables of nodes *1*, *3* and *4*,

$$p(y(4)|y(1), y(3)) = p(y(4)|y(3)) \equiv Y(4) \perp Y(1)|Y(3),$$

where $\perp$ means statistical independence.

This conditional independence structure can also be seen when two nodes have a

common cause. As node *3* influences both nodes *4* and *5*, then:

$$p(y(4)|y(3), y(5)) = p(y(4)|y(3)) \equiv Y(4) \perp Y(5)|Y(3), \text{however } Y(4) \not\perp Y(5).$$

On the other hand, there is no conditional independence when two nodes have the *same effect*, in this case, we have a *v-structure* in the BN. For example, nodes *1* and *2* influence node *3*, then the latter node is known as *collider* and:

$$p(y(1)|y(2), y(3)) \neq p(y(1)|y(3)) \equiv Y(1) \not\perp Y(2)|Y(3), \text{however } Y(1) \perp Y(2).$$

All these conditions above are true because the probability measure P for a directed graph G satisfies the *global directed Markov property*, *i.e.*

$$\mathbf{U} \perp \mathbf{V}|\mathbf{W},$$

if $\mathbf{U}$ is d-separated from $\mathbf{V}$ given $\mathbf{W}$, for all disjoint sets of variables $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ in G (Richardson, 1996).

For DAGs, this global Markov condition is equivalent to the *local directed Markov property*. That is, for any node r in G,

$$Y(r) \perp \mathbf{Y}^{-r}|Pa(r),$$

where $Pa(r)$ is the set of parents of node r and $\mathbf{Y}^{-r}$ is the set of all variables in G, except for node r, its parents and its descendants (Richardson, 1996).

More explicitly, in a BN with nodes represented by the random variables $\mathbf{Y} = (Y(1), \ldots, Y(n))$, the chain rule allows the joint density to be factorized as the product of the distribution of the first node and transition distributions between the following nodes, *i.e.*:

$$p(y(1), y(2), \ldots, y(n)) = p(y(1)) \times \prod_{r=2}^{n} p(y(r)|y(1), \ldots, y(r-1)).$$

Let $Pa(r) \subseteq \{Y(1), \ldots, Y(r-1)\}$ and the Markov properties depicted in the BN states that

a node depends only on its parents. This allows us to simplify the expression above to

$$p(y(1), y(2), \ldots, y(n)) = p(y(1)) \times \prod_{r=2}^{n} p(y(r)|Pa(r)).$$

Therefore, the joint probability of Figure 2.1 can be written as in the following:

$$\begin{aligned} p(y(1), \ldots, y(5)) &= p(y(1)) \times p(y(2)) \times p(y(3)|y(1), y(2)) \times \\ &\quad \times p(y(4)|y(3)) \times p(y(5)|y(3)). \end{aligned}$$

When observed variables are jointly Gaussian, the conditional distribution of variables is defined as $(Y(r)|Pa(r), \boldsymbol{\theta}(r), V(r)) \sim \mathcal{N}(Pa(r)'\boldsymbol{\theta}(r), V(r))$, for $r = 1, \ldots, n$. In this context the regression coefficient $\boldsymbol{\theta}(r)$ represents the functional connectivity strengths (except for intercept).

Several models can be compared based on its joint probability in order to choose the graphical structure that best represents the data. However, some models have the same evidence, *i.e.*, they belong to the same *equivalence class* of models (Friston, 2011). By definition, two network structures are said to be Markov equivalent when they correspond to the same assertions of conditional independence (Heckerman, 1999). For instance, suppose these two graphs: (A) $Y(1) \rightarrow Y(2) \rightarrow Y(3)$ and (B) $Y(1) \leftarrow Y(2) \rightarrow Y(3)$. In both cases, $Y(1)$ and $Y(3)$ are independent given $Y(2)$. In contrast, for this graph: (C) $Y(1) \rightarrow Y(2) \leftarrow Y(3)$, there is another independence structure among nodes, *i.e.*, $Y(1)$ and $Y(3)$ are conditionally dependent. Therefore, graphs (A) and (B) are considered Markov equivalent whilst neither is equivalent to graph (C).

2.4.3 Dynamic Bayesian Network (DBN)

Dynamic Network Models (DNMs) allow multiple networks to be analysed simultaneously or, in other words, analyse the changes in the network over time (Korb and Nicholson, 2004). A classic example in the literature on the use of the DNM is Sampson's monastery study (Sampson, 1968) in which the same network was observed in different intervals of time and, therefore, the evolution of the network was assessed. However, he did not model the dynamic structure expressly as long as he worked with the network in each different time (Goldenberg *et al.*, 2009).

In contrast to the BN, which estimates functional connectivity, the DBN estimates effective connectivity, allowing networks to evolve over time using the GC. Revising notation, now let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_t, \ldots, \mathbf{Y}_T)$ and $\mathbf{Y}_t = (Y_t(1), \ldots, Y_t(n))$, the n-dimension data at time t. The *intra-slice* edges, *e.g.* $Y_t(i) \rightarrow Y_t(j)$, gives the relationships between variables at time t. Assuming that the structure of the BN is the same over time, the relationship among the variables $Y_t(1), \ldots, Y_t(n)$ does not depend on a specific time t. To complete the specification of the model, the *inter-slice* edges must be determined. That is, the interest may be to study change over time in the relationship between the same variable, *e.g.* $Y_{t-1}(i) \rightarrow Y_t(i)$, and different variables, *e.g.* $Y_{t-1}(i) \rightarrow Y_t(j)$.

Under Markov properties, the edges are considered between nodes only at consecutive times, and so the result of one variable at a particular time depends only on what happened at the previous time. Thus, the joint density distribution over $\mathbf{Y}$ is written as:

$$\begin{aligned} p(\mathbf{y}_1, \ldots, \mathbf{y}_T) &= p(\mathbf{y}_1) \times \prod_{t=2}^{T} p(\mathbf{y}_t | \mathbf{y}_{t-1}) \\ &= p(y_1(1)) \times \prod_{r=2}^{n} \left[p(y_1(r) | Pa(r)) \times \prod_{t=2}^{T} p(y_t(r) | Pa(r), \mathbf{y}_{t-1}) \right]. \end{aligned}$$

Chapter 3

The Multiregression Dynamic Model

3.1 Introduction

In the last section, we outlined some of the many types of graphical model used to estimate connectivity. However, many of these models are not concerned with how the brain might work but simply how measurements of brain activity might depend on each other. For instance, for the class of BN models, functional connectivity expressed by the directionality of the edges represents dependence constraints, and not a 'causal' relation in a sense like ones studied by effective connectivity (see Section 2.1). All that might be plausibly concluded is that the absence of an edge might imply that there is no causal relationship between nodes (Queen and Smith, 1993). Thus, to estimate effective connectivity rather than functional connectivity, some models, *e.g.* the DBN, use Granger causality which is defined through specifying certain constraints over time. The GC uses implicitly the idea that the cause must happen *before* an effect. The Multiregression Dynamic Model (MDM) studies the causal relationships between variables in a rather different way. It expresses potential causal hypotheses associated with effective connectivity, through the probabilistic structure over the *contemporaneous* relationships between variables conditional on the past. This is discussed below.

An MDM is a graphical multivariate model for an n-dimensional time series $Y_t(1)$, $Y_t(2)$, ..., $Y_t(n)$, $t = 1, \ldots, T$. Queen and Smith (1993) first built the MDM as a composition

of component univariate regression dynamic linear models (DLMs; West and Harrison, 1997). Each of these components can model smooth changes over time in the parents' effect on a given node during the period of investigation.

We, therefore, begin by describing the regression DLM. Thus consider the relationship between 3 resting-state networks, that is, sets of brain regions that exhibit coherent activity at rest: the Default Mode Network (DMN; node 1), the visual network (node 2), and fronto-parietal network (node 3). One plausible model might be

$$Y_t(1) = \theta_t^{(1)}(1) + \theta_t^{(2)}(1)Y_t(2) + \theta_t^{(3)}(1)Y_t(3) + v_t(1),$$

where $v_t(1)$ is an error term; the parameter $\theta_t^{(1)}(1)$ is the intercept whilst $\theta_t^{(2)}(1)$ and $\theta_t^{(3)}(1)$ are known as connection strengths. Thus, the estimates of regression parameters are found for every time t, allowing the influence of visual and fronto-parietal networks into DMN to vary over time.

The advantage of the formulation above is that because the regression parameters are allowed to be dynamic, the model automatically accounts for some of the variability that might be caused by unobserved variables not recognised in the system (West and Harrison, 1997). For instance, if $Y_t(3)$ is unobserved and evolves slowly and smoothly then it is possible to approximate the true processes on regression only on $Y_t(2)$. Under such an approximation, the intercept parameter, $\theta_t^{(1)}(1)^*$, of the new approximating model can be seen as a function of the missing variable, *i.e.* $\theta_t^{(1)}(1)^* = \theta_t^{(1)}(1) + \theta_t^{(3)}(1)Y_t(3)$, where the unobserved smooth changes in $Y_t(3)$ are modelled by stochastic smooth changes in $\theta_t^{(1)}(1)^*$.

Linear regression models may be seen as a particular case of DLMs, where regression coefficients are fixed over time, and there is an assumption of independent errors. However, this assumption is not appropriate for fMRI data, as they are autocorrelated due to physiological and scanner artifacts.

The usual approach that deals with this problem is for researchers to use a step called prewhitening. Basically, the serial correlation of the time series is removed through an autoregressive model of order 1 (AR(1)). This process allows Ordinary Least Squares to be used for estimation and inference (Poldrack *et al.*, 2011). This preliminary step is not necessary for the DLM, which can explicitly account not only for autocorrelation, but the nonstationarity of the underlying time series. Moreover, the DLM can also deal with change

points or structural breaks in the time series (Petris *et al.*, 2009), in a way we will explore later in this thesis (see Section 3.5.3).

The remainder of this chapter is structured as follows. In Section 3.2, we describe the MDM in details, showing that because of its closed form, the inference process is easily carried out, without using approximate or numerical methods. Then we discuss the use of Bayes factors as a model selection measure for the MDM, and provide two methods to learn the network in Section 3.3. In Section 3.4, we compare the MDM with other methods used to estimate connectivity, highlighting advantages and disadvantages of different approaches. Finally, Section 3.5 gives diagnostic statistics for an MDM.

3.2 The Linear MDM

3.2.1 The Description of the Model

Consider the column vector $\mathbf{Y}_t' = (Y_t(1), \ldots, Y_t(n))$ which denotes the data from n regions at time t. Denote their observed values designated respectively by $\mathbf{y}_t' = (y_t(1), \ldots, y_t(n))$. Let the time series until time t for region $r = 1, \ldots, n$ be $\mathbf{Y}^t(r)' = (Y_1(r), \ldots, Y_t(r))$ and the time series for possible parents of region r at time t be $\mathbf{X}_t(r)' = \{Y_t(1), \ldots, Y_t(r-1)\}$ for $r = 2, \ldots, n$. Note that the n regions in a DAG can always be ordered to ensure that $Pa(r) \subseteq \mathbf{X}_t(r)$, where $Pa(r)$ is the parent set of $Y_t(r)$. The MDM is defined by n observation equations, a system equation and initial information (Queen and Smith, 1993). The observation equations specify the time-varying regression parameters of each region on its parents. The system equation is a multivariate autoregressive model for the evolution of time-varying regression coefficients, and the initial information is given through a prior density for regression coefficients. Thus, the linear multiregression dynamic model is specified in terms of a collection of conditional regression DLMs (West and Harrison, 1997), as follows.

We write the *observation equations* as

$$Y_t(r) = \mathbf{F}_t(r)'\boldsymbol{\theta}_t(r) + v_t(r), \qquad v_t(r) \sim \mathcal{N}(0, V_t(r));$$

where $r = 1, \ldots, n$; $t = 1, \ldots, T$; $\mathcal{N}(\cdot, \cdot)$ is a Gaussian distribution; $\mathbf{F}_t(r)$ is a known function of $Pa(r)$ and is usually defined as $\mathbf{F}_t(r) = \mathbf{M}(r)\mathbf{Y}_t^*$, where $\mathbf{M}(r)$ is $p_r \times (n+1)$ matrix containing only zeros and ones, where ones indicate the parents of $Y_t(r)$, and the first row of

$\mathbf{M}(r)$ is $(1, 0, \ldots, 0)$ representing the intercept; $p_r = |Pa(r)|+1$ counts the number of parents of region r plus one (for the intercept); $\mathbf{Y}_t^* = (1, \mathbf{Y}_t')'$. The observational error, $v_t(r)$, is taken to be independent over t, with variance $V_t(r)$. The p_r-dimensional time-varying regression coefficient is $\boldsymbol{\theta}_t'(r) = (\theta_t^{(1)}(r), \ldots, \theta_t^{(p_r)}(r))$. Generally the parameter $\theta_t^{(1)}(r)$ represents the intercept of the regression of region r whilst $\theta_t^{(i)}(r)$ for $i > 1$ represents the effective connectivity strength for the $(i-1)th$ parent of region r. Concatenating the n regression coefficients as $\boldsymbol{\theta}_t' = (\boldsymbol{\theta}_t'(1), \ldots, \boldsymbol{\theta}_t'(n))$ gives a vector of length $p = \sum_{r=1}^{n} p_r$.

We next write the *system equation* as

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t); \tag{3.1}$$

where $\mathbf{G}_t = \text{blockdiag}\{\mathbf{G}_t(1), \ldots, \mathbf{G}_t(n)\}$, each $\mathbf{G}_t(r)$ being a $p_r \times p_r$ matrix, $\mathbf{w}_t$ is the innovation for the latent regression coefficients, and $\mathbf{W}_t = \text{blockdiag}\{\mathbf{W}_t(1), \ldots, \mathbf{W}_t(n)\}$, each $\mathbf{W}_t(r)$ being a $p_r \times p_r$ matrix. The error $\mathbf{w}_t$ is assumed independent of $\mathbf{v}_s$ for all t and s; $\mathbf{v}_s = (v_s(1), \ldots, v_s(n))$. For most of the development we need only consider $\mathbf{G}_t(r) = \mathbf{I}_{p_r}$, where $\mathbf{I}_{p_r}$ is the p_r-dimensional identity matrix.

For instance, suppose the graphical structure given by Figure 3.1, then the model equations are written as:

$$\begin{aligned}
\boldsymbol{\theta}_t(r) &= \boldsymbol{\theta}_{t-1}(r) + \mathbf{w}_t(r);\ \mathbf{w}_t(r) \sim \mathcal{N}\left(\mathbf{0}, \mathbf{W}_t(r)\right); \\
Y_t(1) &= \theta_t^{(1)}(1) + v_t(1); \\
Y_t(2) &= \theta_t^{(1)}(2) + \theta_t^{(2)}(2)Y_t(1) + v_t(2); \\
Y_t(3) &= \theta_t^{(1)}(3) + \theta_t^{(2)}(3)Y_t(1) + \theta_t^{(3)}(3)Y_t(2) + v_t(3);\ v_t(r) \sim \mathcal{N}\left(0, V_t(r)\right),
\end{aligned}$$

for $r = 1, \ldots, 3$, $p_1 = 1$, $p_2 = 2$ and $p_3 = 3$. The effective connectivity strengths of this example are then $\theta_t^{(2)}(2)$, $\theta_t^{(2)}(3)$ and $\theta_t^{(3)}(3)$.

Finally, the *initial information* is written as

$$(\boldsymbol{\theta}_0|y_0) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0); \tag{3.2}$$

where $\boldsymbol{\theta}_0|y_0$ expresses the prior knowledge of the regression parameters, before observing any data, given the information at time $t = 0$, *i.e.* y_0. The mean vector $\mathbf{m}_0$ is an initial estimate

of the parameters and $\mathbf{C}_0$ is the $p \times p$ variance-covariance matrix. $\mathbf{C}_0$ can be defined as blockdiag$\{\mathbf{C}_0(1), \ldots, \mathbf{C}_0(n)\}$, with each $\mathbf{C}_0(r)$ being a p_r square matrix.

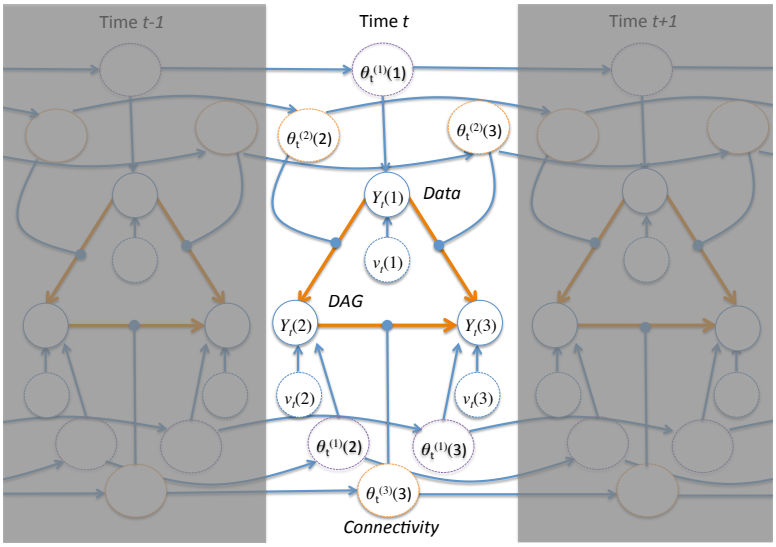


Figure 3.1: Dependence structure for the MDM considering Region 1 as the parent of Region 2 and Region 3; and Region 2 as the parent of Region 3. The solid circles represent observed variables, $Y_t(r)$. The dashed circles represent latent variables: blue for observational errors, $v_t(r)$; violet for the intercept of the regression of Region r, $\theta_t^{(1)}(r)$; $r = 1, 2, 3$; and orange for the effective connectivity strength between two regions, $\theta_t^{(2)}(2)$, $\theta_t^{(2)}(3)$ and $\theta_t^{(3)}(3)$.

There are five important features of this model class discussed in the literature.

1. Although the predictive distributions of each node given its parents are Student t distributed, because the covariates enter the scale function of these conditionals, the joint distribution can be highly non-Gaussian. Queen and Smith (1993) provided examples of this. This feature is useful for fMRI studies, because models that assume that processes are not jointly Gaussian may be better fitted to fMRI data than ones that assume joint Gaussianity;

2. As the values of variables of a particular node and its parents are observed simultaneously, to make predictions, it is necessary to know the marginal forecast distribution for each node in time t, given only the past. This distribution is not generally of a simple form, but it is not hard to calculate its expectation and covariance matrix. Queen and Smith (1993) demonstrated the mean and covariance matrix of the marginal forecast distribution, considering the corrected linear MDM (CLMDM). The linear MDM assumes that the residuals have a Gaussian distribution and the relation between nodes and their parents is linear, where their parents are explanatory variables. In contrast,

the CLMDM uses the residuals of models fitted for parents as regression covariates. The one step ahead mean and covariance matrix of the LMDM were found by Queen *et al.* (2008) and are described in Appendix A. Also, Queen *et al.* (2008) argued the problem that the covariance between root nodes is zero in the LMDM, which sometimes is not expected in a real situation. Therefore they proposed to include in the model a set of variables that explain the correlation between roots as a parent of them;

3. Each LMDM is defined in part by a directed acyclic graph (DAG) whose vertices are observed fMRI series at a given time. In addition, its directed edges represent the existence of a dependence on those contemporaneous observations that are explicitly included as regressors to the receiving variable. In our context, therefore, these directed edges denote the hypothesis that direct contemporaneous relationships might exist between a variable and its parents. The directionality of the edges can be interpreted as being 'causal' in a sense that is carefully argued in Queen and Albers (2009);

4. Dependence relationships between each component and its contemporaneous parents — as represented by the corresponding regression coefficients — are allowed to drift with time. Therefore, unlike a static BN, the MDM *models* dynamic links and so allows us to discriminate between models that would be Markov equivalent in their static versions. Queen and Albers (2009) showed this result using real traffic flows data. We will also discuss this question considering different sample sizes and dynamic levels using synthetic data in Chapter 4;

5. The class of MDM can be further modified to include other features that might be necessary in a straightforward and convenient manner. For instance, Queen and Albers (2009) showed that a causal relationship could be better identified using the intervention process. In addition, Anacleto Junior *et al.* (2013a) worked with heteroscedasticity and measurement errors in the LMDM. Yet Anacleto Junior *et al.* (2013b) dealt with cycle problems in the time series using cubic splines in the MDM. Some methods used to check and to embellish the MDM are discussed in Section 3.5.

3.2.2 The Inferential Process

When the observational variances are unknown and constant, *i.e.* $V_t(r) = V(r)$ for all t, by defining $\phi(r) = V(r)^{-1}$, a prior

$$(\phi(r)|y_0) \sim \mathcal{G}\left(\frac{n_0(r)}{2}, \frac{d_0(r)}{2}\right), \tag{3.3}$$

where $\mathcal{G}(\cdot,\cdot)$ denotes a Gamma distribution, leads to a conjugate analysis where conditionally each component of the marginal likelihood has a Student t distribution. In order to use this conjugate analysis it is convenient to reparameterise the model as $\mathbf{W}_t(r) = V(r)\mathbf{W}_t^*(r)$ and $\mathbf{C}_0(r) = V(r)\mathbf{C}_0^*(r)$. For a fixed innovation signal matrix $\mathbf{W}_t^*(r)$ this change implies no loss of generality (West and Harrison, 1997).

When the estimation process is performed using only the data observed thus far, *i.e.* $\mathbf{Y}^t$, we call it a filtering approach; when the estimation is performed using all data $\mathbf{Y}^T$, we call it a smoothing approach. Filtered estimation is suitable when data are available sequentially in time, as in financial applications, when a certain rate needs to be estimated every day. Moreover, it is used in the calculation of the predictive likelihood and, at the final time point, for model selection. Smoothed estimation provides an understanding of the complete series after it has been observed in a certain period, and so the interest is to answer the question "What happened?" (Petris *et al.*, 2009).

Queen and Smith (1993) showed that when the parameters are mutually independent at $t = 0$ for each variable, which happens when $\mathbf{C}_0$ is set to be block diagonal, then

$$\perp_{r=1}^{n} \boldsymbol{\theta}_t(r)|\mathbf{y}^t \text{ and}$$
$$\boldsymbol{\theta}_t(r) \perp \mathbf{Y}^t(r+1), \mathbf{Y}^t(r+2), \ldots, \mathbf{Y}^t(n)|\mathbf{y}^t(1), \ldots, \mathbf{y}^t(r).$$

Therefore, the posterior filtered distributions are found recursively as shown in Appendix A, *i.e.*

$$\begin{aligned}
(\boldsymbol{\theta}_t(r)|\mathbf{y}^{t-1}(r), \phi(r)) &\sim \mathcal{N}\left(\mathbf{m}_{t-1}(r), \mathbf{C}_{t-1}^*(r)\phi(r)^{-1}\right), \\
(\boldsymbol{\theta}_t(r)|\mathbf{y}^t) &\sim \mathcal{T}_{n_t(r)}(\mathbf{m}_t(r), \mathbf{C}_t(r)) \text{ and} \\
(\phi(r)|\mathbf{y}^t) &\sim \mathcal{G}\left(\frac{n_t(r)}{2}, \frac{d_t(r)}{2}\right),
\end{aligned} \tag{3.4}$$

where $t = 1, \ldots, T$, $\mathcal{T}_{n_t(r)}(\cdot, \cdot)$ is a noncentral t distribution with $n_t(r)$ degrees of freedom, and

$$\begin{aligned}
\mathbf{m}_t(r) &= \mathbf{m}_{t-1}(r) + \mathbf{A}_t(r)e_t(r); \\
\mathbf{A}_t(r) &= \mathbf{R}_t^*(r)\mathbf{F}_t(r)/Q_t^*(r); \\
\mathbf{R}_t^*(r) &= \mathbf{C}_{t-1}^*(r) + \mathbf{W}_t^*(r); \\
Q_t^*(r) &= 1 + \mathbf{F}_t'(r)\mathbf{R}_t^*(r)\mathbf{F}_t(r); \\
e_t(r) &= Y_t(r) - f_t(r); \\
f_t(r) &= \mathbf{F}_t'(r)\mathbf{m}_{t-1}(r); \\
\mathbf{C}_t^*(r) &= \mathbf{R}_t^*(r) - \mathbf{A}_t(r)\mathbf{A}_t'(r)Q_t^*(r); \\
\mathbf{C}_t(r) &= S_t(r)\mathbf{C}_t^*(r) = [d_t(r)/n_t(r)]\mathbf{C}_t^*(r); \\
n_t(r) &= n_{t-1}(r) + 1; \\
d_t(r) &= d_{t-1}(r) + e_t(r)^2/Q_t^*(r).
\end{aligned}$$

When $\mathbf{W}_t^*$ is unknown, the reparameterised model simplifies the analysis and allows us to define the innovation signal matrix indirectly in terms of a single hyperparameter for each component DLM called a *discount factor* (West and Harrison, 1997; Petris *et al.*, 2009), especially for model selection purposes we have used here. This vastly reduces the dimensionality of the model class whilst in practice often loses very little in the quality of fit. This well used technique expresses different values of $\mathbf{W}_t^*$ in terms of the loss of information in the change in $\boldsymbol{\theta}$ between times $t-1$ and t. More precisely, by equations (3.1) and (3.4) the prior distribution of $\boldsymbol{\theta}_t(\mathrm{r})$ is

$$\begin{aligned}
(\boldsymbol{\theta}_t(r)|\mathbf{y}^{t-1}(r), \phi(r)) &\sim \mathcal{N}\left(\mathbf{m}_{t-1}(r), \mathbf{C}_{t-1}^*(r)\phi(r)^{-1}\right) + \mathcal{N}\left(\mathbf{0}, \mathbf{W}_t^*(r)\phi(r)^{-1}\right) \\
&\sim \mathcal{N}\left(\mathbf{m}_{t-1}(r), \mathbf{R}_t^*(r)\phi(r)^{-1}\right).
\end{aligned}$$

Therefore, as the variance of innovation residuals ($\mathbf{W}_t^*(r)$) is unknown, if we can assume that the prior variance at time t, which consists of $\mathbf{R}_t^*(r)$, is well approximated by a percentage of the posterior variance at time $t-1$, which consists of $\mathbf{C}_{t-1}^*(r)$, then

$\mathbf{R}_t^*(r) = \mathbf{C}_{t-1}^*(r)/\delta(r)$ for some $\delta(r) \in (0,1]$, and we have a similar expression for

$$\mathbf{W}_t^*(r) = \frac{1-\delta(r)}{\delta(r)}\mathbf{C}_{t-1}^*(r).$$

Thus $\mathbf{W}_t^*(r)$ is defined deterministically through $\delta(r)$ as a discounted value of $\mathbf{C}_{t-1}^*(r)$. Note that when $\delta(r) = 1$, $\mathbf{W}_t^*(r) = 0\mathbf{I}_{p_r}$, there are no stochastic changes in the state vector and we degenerate to a conventional standard multivariate Gaussian prior to posterior analysis. For any choice of discount factor $\delta(r)$ and any MDM the recurrences given above provide a closed form expression for this marginal likelihood. This means that we can estimate $\delta(r)$ simply by maximising this marginal likelihood, performing a direct one-dimensional optimisation over $\delta(r)$, analogous to that used in Heard *et al.* (2006) to complete the search algorithm. The selected component model is then the one with the discount factor giving the highest associated Bayes factor score, as we will see later.

The smoothed estimation of the parameters for each variable follows a retrospective analysis, starting with $t = T-1$ and continues until $t = 1$, via (see demonstration in Appendix A)

$$(\boldsymbol{\theta}_t(r)|\mathbf{y}^T) \;\sim\; \mathcal{T}_{n_T(r)}\left(\mathbf{sm}_t(r), \mathbf{sC}_t(r)\right),$$

where

$$\begin{aligned}
\mathbf{sm}_t(r) &= \mathbf{m}_t(r) + \mathbf{C}_t^*(r)(\mathbf{R}_{t+1}^*(r))^{-1}(\mathbf{sm}_{t+1}(r) - \mathbf{m}_t(r));\\
\mathbf{sC}_t^*(r) &= \left[\mathbf{C}_t^*(r) - \mathbf{C}_t^*(r)(\mathbf{R}_{t+1}^*(r))^{-1}(\mathbf{R}_{t+1}^*(r) - \mathbf{sC}_{t+1}^*(r))(\mathbf{R}_{t+1}^*(r))^{-1}\mathbf{C}_t^*(r)\right]\phi^{-1}(r);\\
\mathbf{sC}_t(r) &= S_T(r)\mathbf{sC}_t^*(r).
\end{aligned}$$

The conditional forecast distribution of $(Y_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t(r))$ given the past is also identical to the DLM (see Appendix A), *i.e.*:

$$(Y_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t(r)) \sim \mathcal{T}_{n_{t-1}(r)}(f_t(r), Q_t(r)), \tag{3.5}$$

where

$$\begin{aligned} f_t(r) &= \mathbf{F}_t'(r)\mathbf{m}_{t-1}(r) \\ Q_t(r) &= \frac{d_{t-1}(r)}{n_{t-1}(r)}[\mathbf{F}_t'(r)\mathbf{R}_t^*(r)\mathbf{F}_t(r)+1]. \end{aligned}$$

The joint density over the vector of observations associated with any MDM series can be factorized into the product of the density of the first node and the (conditional) transition densities between the subsequent nodes (Queen and Smith, 1993). The joint log predictive likelihood (LPL) is then calculated based on (3.5) as

$$\begin{aligned} \text{LPL} &= \log p(\mathbf{y}) \\ &= \sum_{r=1}^{n} \log p(\mathbf{y}(r)|\mathbf{x}(r)) \\ &= \sum_{r=1}^{n}\sum_{t=1}^{T} \log p(y_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t(r)), \end{aligned} \tag{3.6}$$

where $\mathbf{x}_t(1)$ is empty.

Although $\mathbf{F}_t$ is assumed fixed in the DLM, $\mathbf{F}_t$ is a function of $\mathbf{X}_t(r)$ in the MDM, and therefore, it is also a random variable. Moreover, $Y_t(r)$ and $\mathbf{X}_t(r)$ are observed simultaneously; thus, it may be necessary to know the marginal forecast distribution for each $Y_t(r)$, given only the past $\mathbf{Y}^{t-1}$. The one step ahead mean and covariance matrix of the LMDM were found by Queen *et al.* (2008) and are described in Appendix A.

3.2.3 Priors

Two priors are considered in the MDM: one in the process of model selection (*e.g.* learning network) and other in the estimation of parameters. They are called respectively by *the model prior* and *the parameter priors* (Heckerman, 1999). The former will be discussed in Section 3.2.4 whilst the latter was firstly shown in the previous section, *i.e.* the prior for regression parameters is given in equation (3.2) and, when the observational variance is unknown, the prior for the observational precision is given in equation (3.3).

Now we will show the impact of the parameter priors on the inference process. Consider, for example, the connectivity between Regions 3 and 4 in Figure 3.2. Under the observation equation: $Y_t(4) = \theta_t Y_t(3) + v_t(4)$, the conditional forecast mean of the variable

$Y_{t+1}(4)$ is $m_t(4)Y_{t+1}(3)$, as in equation (3.5), and from equation (3.4), $m_t(4)$ can be rewritten as

$$\begin{aligned} m_t(4) &= A_t(4)Y_t(4) + m_{t-1}(4)(1 - A_t(4)Y_t(3)), \\ \text{where } A_t(4) &= \frac{R_t^*(4)Y_t(3)}{R_t^*(4)Y_t(3)^2 + 1}. \end{aligned}$$

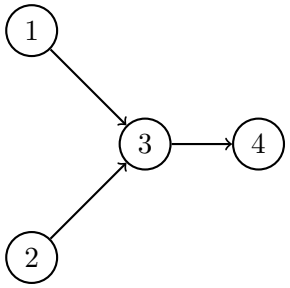


Figure 3.2: A graphical structure considering 4 nodes.

Thus the calculation of the current forecast mean is based on $(1 - A_t(4)Y_t(3))\%$ of the previous mean m_{t-1}. When the latter is replaced by its own equation in the function of $m_{t-2}(4)$, we find

$$\begin{aligned} m_t(4) &= A_t(4)Y_t(4) + A_{t-1}(4)Y_{t-1}(4)(1 - A_t(4)Y_t(3)) + \\ &+ m_{t-2}(4)(1 - A_{t-1}(4)Y_{t-1}(3))(1 - A_t(4)Y_t(3)), \end{aligned}$$

and then the forecast mean of the second previous time m_{t-2} contributes $(1-A_{t-1}(4)Y_{t-1}(3))(1-A_t(4)Y_t(3))\%$ to current forecast mean. Following the same reasoning for $m_{t-2}(4)$ onwards, we find the forecast mean as a function of the prior mean as

$$\begin{aligned} m_t(4) &= A_t(4)Y_t(4) + \sum_{k=1}^{t-1}[A_k(4)Y_k(4) \prod_{j=k+1}^{t} (1 - A_j(4)Y_j(3))] + \\ &+ m_0(4)\prod_{i=1}^{t}(1 - A_i(4)Y_i(3)). \end{aligned} \tag{3.7}$$

Therefore note that as t increases the value of $\prod_{i=1}^{t}(1 - A_i(4)y_i(3))$ decays to zero and so the importance of the prior mean $m_0(4)$ in the calculation of $m_t(4)$ decreases.

We studied the impact of the prior distribution in the calculation of the posterior distribution of a regression parameter using real fMRI data. This dataset consists of 176

time points and 36 subjects (see more detail about this data in Section 4.3 and Smith *et al.*, 2009), and we considered node 1, visual region, as parent of node 2, DMN. The inferential process was led using the values of 0 and 1 for the hyperparameter $m_0(2)$ and the values of 0.5, 1 and 3 for the hyperparameter $C_0^*(2)$. Figure 3.3 (*left*) shows the average of the contribution of prior mean $m_0(2)$ in the calculation of posterior mean $m_t(2)$ (as in the equation (3.7)) over 36 subjects. Note that this contribution is less than 1% from time 17 for all values of prior hyperparameters (a similar result can be seen for the constant model in West and Harrison, 1997, chapter 2). The centre picture shows the posterior mean $m_t(2)$ and the right picture shows the posterior variance $C_t(2)$ for a particular subject with the same values of hyperparameters. In general, the average of difference between the results of the prior hyperparameters over subjects is less than 0.02 for the posterior mean from time 11 and less than 0.002 for the posterior variance from time 12. Therefore, after time 10 the posterior distribution is almost the same regardless of the typical values we might choose for the hyperparameters of the prior (see, in the centre and right pictures, that the different colour lines become almost the only one after the point 10). In the next section, we present a model selection criteria whilst in Section 3.5.1 we describe how we have nevertheless matched priors to minimize this small effect in the consequent Bayes factor scores driving the model selection.

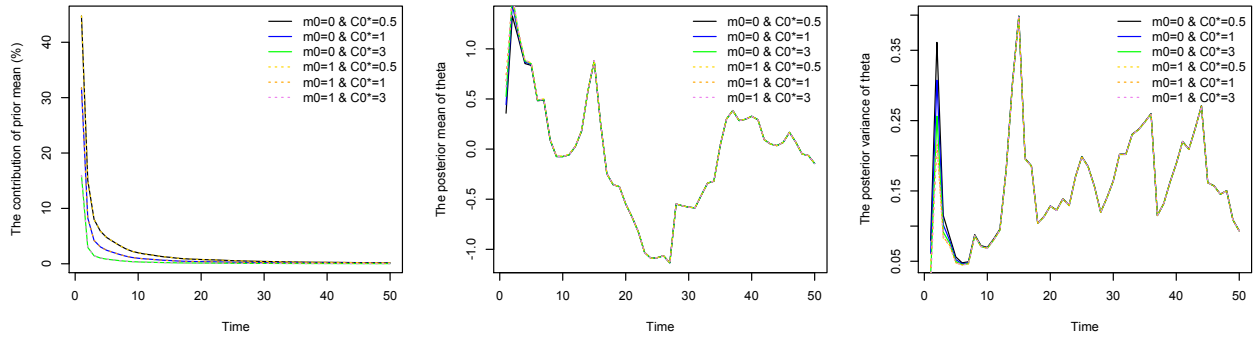


Figure 3.3: In this picture we show the impact of priors in the posterior distribution of the connectivity $Y(1) \rightarrow Y(2)$, where node 1 is visual region and node 2 is DMN. The left picture shows the average of the contribution of the prior mean $m_0(2)$ in the calculation of posterior mean $m_t(2)$, defined as $\prod_{i=1}^{t}(1 - A_i(2)y_i(2))\%$, over 36 subjects, by different values of hyperparameters. They are less than 1% from time 17. The center picture shows the posterior mean $m_t(2)$ whilst the right picture shows the posterior variance $C_t(2)$ for subject 19 by the same values of hyperparameters. See text for more details.

3.2.4 Criteria for Model Selection

Suppose we want to compare $\mathcal{M}$ different models. Let $p(M|\mathbf{y}^T)$ be the posterior probability of model $M \in \{1, \ldots, \mathcal{M}\}$ defined as

$$p(M|\mathbf{y}^T) \propto p(\mathbf{y}^T|M)p(M),$$

where $p(M)$ is the *model prior* and $p(\mathbf{y}^T|M)$ is the predictive likelihood given by equation (3.6) in log scale. If all structures are a priori equally likely, *i.e.* $p(M_1) = p(M_2) = \ldots = p(M_{\mathcal{M}})$, where M_i is shorthand for event $\{M = i\}$, then the comparison criteria is now based on the predictive distributions. For instance, if we compare two models, then

$$\frac{p(M_1|\mathbf{y}^T)}{p(M_2|\mathbf{y}^T)} = \frac{p(\mathbf{y}^T|M_1)p(M_1)}{p(\mathbf{y}^T|M_2)p(M_2)} = \frac{p(\mathbf{y}^T|M_1)}{p(\mathbf{y}^T|M_2)}.$$

This ratio of two predictive likelihoods is called the Bayes factor (BF; see *e.g.* Jeffreys, 1961 and Gamerman, 1997; in the context of state space model: Fruhwirth-Schnatter, 1995 and West and Harrison, 1997). Therefore, the Bayes factor on the log scale (logBF) is defined as

$$\text{logBF} = \text{LPL}(M_1) - \text{LPL}(M_2).$$

West and Harrison (1997) suggested a criterion of ± 1 for the logBF; that is, logBF≥ 1 is evidence for model 1, while logBF≤ -1 is evidence for model 2. If $-1 <$logBF< 1, the evidence is equivocal.

3.3 The Process of Search Networks Applied to the MDM

It is well known that finding the highest scoring model even within the class of vanilla BNs is challenging. Even after using prior information to limit this number to scientifically plausible ones, it is usually necessary to use search algorithms to guide the selection. However recently there have been significant advances in performing this task (see *e.g.* Spirtes *et al.*, 2000; Meek, 1997; Ramsey *et al.*, 2010; Cussens, 2010; Cowell, 2013), and below we make use of one of the most powerful methods currently available. We exploit the additive nature of the

MDM score function — equation (3.6), where there are exactly n terms, one per region. Each region has 2^{n-1} possible configurations, according to whether each other region is included or excluded as a parent. Thus, exhaustive computation of all possible score *components* is feasible; for example, a 10 node network has only 5,120 possible components. Using the constrained integer programming method described below, we have a method that allows the selection of the optimal MDM with only modest computational effort. However, because of the fully Bayesian formulation of the processes, it is also possible to adapt established predictive diagnostics to this domain to further examine the discrepancies associated with the fit of the best scoring model and adjust the family where necessary. How this can be done is explained in Section 3.5, and the results of such an analysis are illustrated in Chapter 4.

Model selection algorithms for probabilistic graphical models can be classified into two categories: the *constraint-based method* and the *search-and-score method*. The former uses the conditional independence constraints whilst the latter chooses a model structure that provides the best trade-off between the fit to data and model complexity using a scoring metric. For instance, the PC-algorithm is a constraint-based method and searches for a partially directed acyclic graph (PDAG) (Spirtes *et al.*, 2000; Meek, 1995; Kalisch and Bühlmann, 2008). A PDAG is a graph that may have both undirected and directed edges but no cycles. This algorithm assumes that the random variables follow a multivariate Gaussian distribution and then uses the partial correlation to infer conditional independencies. It searches for a PDAG that represents a Markov equivalence class, beginning with a complete undirected graph. Then, edges are gradually deleted according to discovered conditional independence. This means that edges are firstly deleted if they link variables that are unconditionally independent. The same applies if the variables are independent conditional on one other variable, conditional on two other variables, and so on. Figure 3.4 shows how this algorithm works using one example (Spirtes *et al.*, 2000). The first row of this figure displays the true graph in which data were generated (on the left side), and the complete undirected graph which was assumed initially by PC-algorithm (on the right side). Then, for each pair of two nodes, it was checked if its observed variables were independent. But, as all variables were actually unconditional dependents (see true graph), no edge was removed in this second step. Following, the edges were deleted according to conditional independence

on one and two variables, as shown in the third and the fourth row of this figure. Finally, the directed edges $C \rightarrow E$ and $D \rightarrow E$ were defined as long as C and D were not independent given E, and so there was a collision in this latter node. Therefore, the result of PC-algorithm in this example was a PDAG shown in the bottom of Figure 3.4.

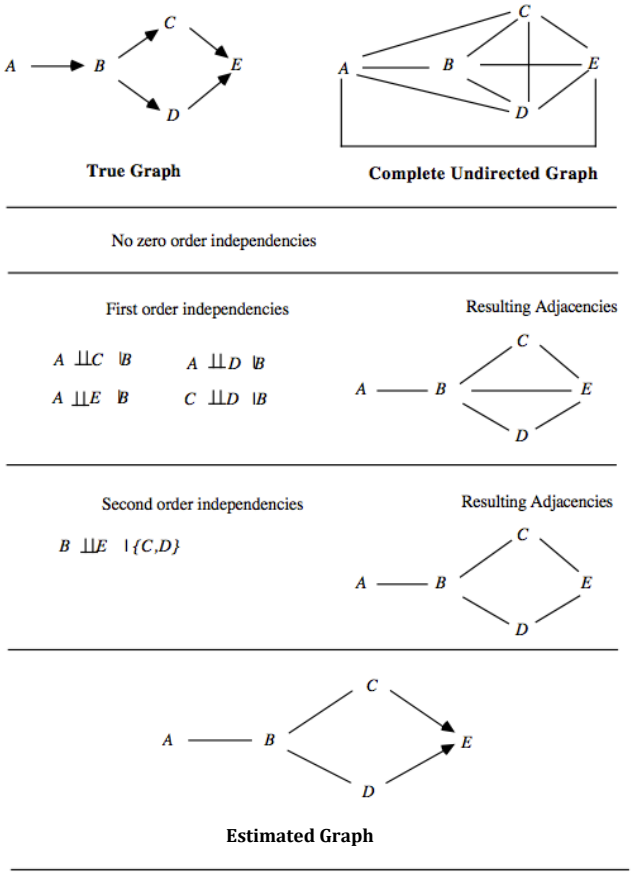


Figure 3.4: An example for PC-algorithm. (Figure from Spirtes *et al.*, 2000, Chapter 5).

On the other hand, the Greedy Equivalence Search (GES) is a search-and-score method using the Bayes Information Criterion (BIC; Schwarz, 1978) to score the candidate structures (Chickering, 2002; Meek, 1997). As PC, GES also searches for PDAG and assumes Gaussian distribution. The algorithm starts with an empty graph, in which all nodes are independent and then gradually, all possible single-edges are compared, and one is added each time. This process stops when the BIC score no longer improves. At this point, the reverse process is then driven in which edges are removed in the way described above. Again, when the improvement of the score is not possible, the graphical structure that represents a DAG equivalence class is chosen (Ramsey *et al.*, 2010). Note that as PC and GES search for a Markov equivalence class, it is not possible to use them with the MDM,

which discriminates graphical structures that belong to the same equivalence class.

3.3.1 Scoring the MDM Using an Integer Programming Algorithm

The insight we use here is that the problem of searching graphical structures for MDM can be seen as an optimization problem suitable for solving with an *integer programming* (IP) algorithm. We use IP for the first time to search for the graphical structure for the MDM, adapting established IP methods used for BN learning. Cussens (2010) developed a search approach for the BN pedigree reconstruction with the help of auxiliary integer-valued variables, whilst Cowell (2013) used a dynamic programming approach with a greedy search setting for this same problem. Jaakkola *et al.* (2010) also applied IP, but instead of using auxiliary variables, they worked with *cluster-based constraints* as explained below. Cussens (2011) took a similar approach to Jaakkola *et al.* but with different search algorithms. The MDM-IP algorithm follows the approach provided by Cussens (2011), but with the MDM scores given by LPL rather than BN scores, as we will show below.

An Integer Programming algorithm is a search-and-score method and a *standard* form of IP is defined as the problem of maximising $\mathbf{c}'\mathbf{x}$, $\mathbf{x}$ being an integer and with the constraints of $\mathbf{Ax} \leqslant \mathbf{b}$ and $\mathbf{x} \geqslant \mathbf{0}$ (Williams, 2009). The expression to be maximised is named *objective function*. For instance, suppose the problem of maximizing x_2 with the linear constraints of $x_1 - x_2 \leqslant 0$ and $6x_1 + x_2 \leqslant 18$, considering yet x_1 and x_2 as integer and non-negative values. Figure 3.5 shows the problem. The blue line represents the equation $x_1 - x_2 = 0$ whilst the orange line represents $6x_1 + x_2 = 18$. The integer values of x_2 (green points) inside the green triangle satisfy the constraints and, therefore, the optimal solution of this IP is $x_2 = 2$. However, if the variables are considered non-negative and belong to the set of real numbers, the objective value of the relaxation is 2.6 (violet point).

Our IP problem consists of finding a graphical structure which maximizes the joint predictive likelihood with the constraints of a DAG. Recall that we use the joint log predictive likelihood (LPL) to score candidate models and that this likelihood has a closed form as a product of Student t-distributions. Therefore, for any candidate model m, LPL(m) is a sum of n 'local scores', one for each node r, and the local score for $Y_t(r)$ is determined by the choice of parent set $Pa_m(r)$ specified by the model m. Let $c(r, Pa_m(r))$ denote this local score, so that $\text{LPL}(m) = \sum_{r=1}^{n} c(r, Pa_m(r))$.

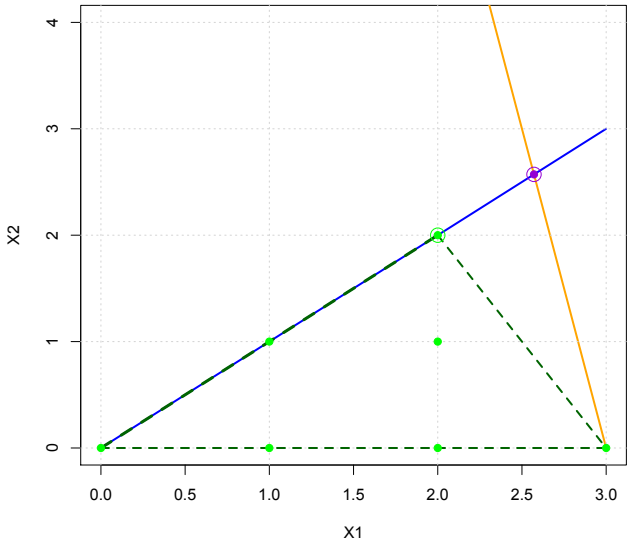


Figure 3.5: An example of integer programming and its linear programming relaxation.

Rather than viewing model selection for the MDM directly as a search for a model m, we view it as a search for n subsets $Pa(1), \ldots Pa(n)$ which maximise $\sum_{r=1}^{n} c(r, Pa(r))$ subject to there existing an MDM model m with $Pa(r) = Pa_m(r)$ for $r = 1, \ldots, n$. We thus choose to see model selection as a problem of constrained discrete optimisation. In the first step of our approach we compute local scores $c(r, Pa)$ for all possible values of Pa and r, where Pa may be $\emptyset$. Next we create indicator variables $I(r \leftarrow Pa)$, one for each local score. $I(r \leftarrow Pa) = 1$ indicates that $Pa_m(r) = Pa$ in some candidate model m. Note that creating all these local scores and variables is practical considering the number of nodes in this application. The model selection problem can now be posed in terms of the $I(r \leftarrow Pa)$ variables:

Choose values for the $I(r \leftarrow Pa)$ variables to maximise

$$\sum_{r} c(r, Pa) I(r \leftarrow Pa) \tag{3.8}$$

subject to there existing an MDM model m with $I(r \leftarrow Pa) = 1$ iff $Pa = Pa_m(r)$.

We choose an IP representation for this problem. To be an IP problem the objective function must be linear, and all variables must take integer values. Both of these are indeed the case in this application. However, in addition, all constraints on solutions must be linear—an issue that we now consider.

Clearly, any model m determines exactly one parent set for each $Y_t(r)$. This is represented by the following n linear *convexity constraints*:

$$\forall r = 1, \ldots n : \sum_{Pa} I(r \leftarrow Pa) = 1. \tag{3.9}$$

It is not difficult to see that constraints (3.9) alone are enough to ensure that any solution to our IP problem represents a directed graph (*digraph*). Additional constraints are required to ensure that any such graph is *acylic*.

There are a number of ways of ruling out cyclic digraphs. We have found the most efficient method is to use *cluster constraints* first introduced by Jaakkola *et al* (2010). These constraints state that in an acyclic digraph any subset ('cluster') of vertices must have at least one member with no parents in that subset. Formally:

$$\forall C \subseteq \{1, \ldots, n\} : \sum_{r \in C} \sum_{Pa : Pa \cap C = \emptyset} I(r \leftarrow Pa) \geq 1. \tag{3.10}$$

Maximising the linear function (3.8) subject to linear constraints (3.9) and (3.10) is an IP problem. To solve our IP problem we have used the GOBNILP system (Cussens, 2011; Bartlett and Cussens, 2013). In GOBNILP the convexity constraints are present initially but not the cluster constraints. As is typical in IP solving, GOBNILP first solves the *linear relaxation* of the IP where the $I(r \leftarrow Pa)$ variables are allowed to take any value in $[0, 1]$ not just 0 or 1. The linear relaxation can be solved very quickly. GOBNILP then searches for cluster constraints (3.10) which are violated by the solution to the linear relaxation. Any such cluster constraints are added to the IP (as so-called *cutting planes*) and the linear relaxation of this new IP is then solved and cutting planes for this new linear relaxation are then sought, and so on. If at any point the solution to the linear relaxation represents an acylic digraph, the problem is solved. In all cases, we are able to solve the problem to optimality, returning an MDM model which is guaranteed to have maximal joint log predictive likelihood (LPL).

To sum up, the idea of this search method is that the algorithm begins considering the convexity constraints, looking for the best-scoring parent set for each node independently. Then, the cluster constraint is verified using the equation (3.10), and if it is violated, the node that has the worst-scoring parent set among all nodes that form the cycle is selected. Then, the next best-scoring parent set for this selected node is found. The cluster constraint

is tested again for this new graph, and if it is a DAG, the algorithm finishes, otherwise the algorithm continues as described above.

For instance, consider a search problem for 3 variables. Table 3.1 shows the local scores $c(r, Pa)$ for all possible values of $Pa(r)$ for $r = 1, \ldots, 3$. As said before, the first step of the IP algorithm is

1. to maximise the objective function regarding the convexity constraints, *i.e.* there is only one set of parents for each node. Therefore, the best scoring model consists of node 1 with no parents, nodes 1 and 2 as the parent of node 3, and nodes 1 and 3 as the parent of node 2, see Figure 3.6(a).

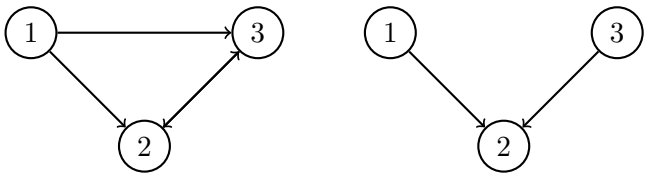
2. The cluster constraint is then tested. However, it is a cyclic graph, because of the cluster formed by nodes 2 and 3.

3. The following step is then to compare the scores of these two nodes. Indeed node 3 has the lower score for this parent set than node 2.

4. Then the next best-scoring parent set for node 3 is node 2 as its parent. However, again there is a cycle between nodes 2 and 3, and the score for node 3 is lower than for node 2.

5. The next best-scoring parent set for node 3 is no parent for this node.

6. Finally, it is a DAG, as shown in Figure 3.6(b).



(a) (b)

Figure 3.6: The IP solution regarding the scores of table 3.1, the convexity constraints (a) and the cluster constraints (b).

3.3.2 Directed Graph Model Search

Bidirectional communication between some brain regions is often expected. Thus, cyclic graphs (*i.e.* graphs allowing cycles) may better represent brain networks than DAGs. Therefore, we are also considering search for graphical structure without the constraints of DAG

Node	Parent	Score
1	No	-1469
	2	-1567
	3	-1646
	2 and 3	-1655
2	No	-1169
	1	-1140
	3	-1110
	1 and 3	-997
3	No	-1119
	1	-1193
	2	-1060
	1 and 2	-1056

Table 3.1: Evidence for each node under all possible sets of parents. The higher the score, the higher evidence for this particular model.

(*cluster constraints*). Because the predictive likelihood factors by node, this reduces to choosing the set of parents that maximise the LPL for each node independently. The main problem with this class is that the composite model typically will not correspond to a single probability model. The output is, therefore, a simple heuristic. It is nevertheless very useful as an additional exploratory data analysis tool.

This approach is called as the *MDM-DGM algorithm* (DGM is short for Directed Graph Model). The analysis of cyclic graphs is unlike the analysis of DAGs in some aspects. Spirtes *et al.* (2000, chapter 12) compared some properties such as the Markov condition and factorizability between DAG and cyclic graphs. For instance, consider this directed cyclic graph (DCG): $1 \rightarrow 2 \rightleftarrows 3$. As shown in Section 2.4.2, the DAG satisfies the local Markov property, *i.e.* the variable of node i, given its parents, is independent of all other variables, except for its parents and descendants. But, DCG does not always satisfy this property, as $\mathbf{Y}(3)$ is not independent of $\mathbf{Y}(1)$ given $\mathbf{Y}(2)$ — node 2 is the parent of node 3, because the path $1 \rightarrow 2 \leftarrow 3$.

Considering now an other example, the DCG in Figure 3.7, as each variable is generated from each other, they are dependent. However, note that, in an associated undirected graph (changing all directed edges by undirected ones), $\mathbf{Y}(1)$ is conditional independent of $\mathbf{Y}(3)$ given $\mathbf{Y}(2)$ and $\mathbf{Y}(4)$ and, in the same way, $\mathbf{Y}(2)$ and $\mathbf{Y}(4)$ are conditional independent given $\mathbf{Y}(1)$ and $\mathbf{Y}(3)$. Therefore, Spirtes (1995) asserted that the d-separation is informative in cyclic graphs. That is, if two nodes, say 1 and 3, are d-separated by a third node, say 2,

then the partial correlation of $\mathbf{Y}(1)$ and $\mathbf{Y}(3)$ given $\mathbf{Y}(2)$ becomes null. Spirtes *et al.* (2000) showed that the global Markov property for directed (acyclic or cyclic) graph holds for the linear structural equation model. That is, a distribution P is represented by a directed graph G if and only if whenever **I** and **II** are d-separated given **III** in G, where **I**, **II** and **III** are disjoint sets of nodes in G, the two sets of variables of nodes **I** and **II** are conditional independent in P given the variables of **III**.

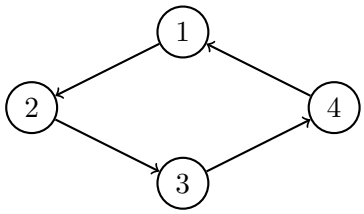


Figure 3.7: An example of cyclic graph.

Another aspect that differs between cyclic and acyclic graphs is factorizability. In DAGs, the joint distribution of variables is defined as the product of the conditional distribution of each variable given its parents, whilst it may be not possible in DCG. For instance, suppose that $1 \rightleftarrows 2$, if $p(\mathbf{y}(1), \mathbf{y}(2)) = p(\mathbf{y}(1)|\mathbf{y}(2))p(\mathbf{y}(2)|\mathbf{y}(1))$, then it means that $\mathbf{Y}(1)$ is independent of $\mathbf{Y}(2)$, as long as

$$\begin{aligned} p(\mathbf{y}(1), \mathbf{y}(2)) &= p(\mathbf{y}(1)|\mathbf{y}(2))p(\mathbf{y}(2)|\mathbf{y}(1)) \\ p(\mathbf{y}(1)|\mathbf{y}(2))p(\mathbf{y}(2)) &= p(\mathbf{y}(1)|\mathbf{y}(2))p(\mathbf{y}(2)|\mathbf{y}(1)) \\ p(\mathbf{y}(2)) &= p(\mathbf{y}(2)|\mathbf{y}(1)), \end{aligned}$$

which does not represent the dependence constraint asserted by the graph. The joint model can no longer be guaranteed to be Gaussian and the analysis provided by the calculation above no longer leads to a Bayesian conjugative analysis. In fact in the non-stochastic case we degenerate to an SEM model, see Section 2.3, which are notoriously hard formally to estimate. In this sense, the DGM can be seen as a class of a structurally dynamic SEMs (see *e.g.* Koster, 1996). So this emphasises that the DGM models we fit here simply provide a heuristic, summarising the best features of classes of MDM and not a fully and unambiguously specified model in itself.

3.3.3 The Running Time of the MDM-IPA and the MDM-DGM

The learning network process follows two steps. Initially, the scores for each set of parents for individual nodes are found. Then the MDM-IPA (or the MDM-DGM) is applied to discover the best MDM over all nodes.

Step I - calculating the scores

The run-time of the first step (finding the scores) of course depends critically on the number of nodes and the sample size. It is necessary to fit a linear dynamic model for every node and every set of parents — there are 2^{n-1} possible sets of parents per node (see *e.g.* Table 3.1). In addition, when the innovation variance is unknown, it is necessary to fit every model several times, according to different values of the discount factor. Then the model (with a particular value of DF) which provides the highest score is selected.

Table 3.2 shows the time taken in minutes to find the scores for different numbers of nodes and sample sizes, on a 2.7 GHz quad-core Intel Core i7 linux host with 16 GB, using the software R[1]. The discount factor was chosen in the range from 0.5 to 1.0 with an increments of 0.01. There is a sharp increase in the process time when the underlying graph has 10 or more nodes. As future work, we expect to change the program using the *C language*[2] and procedures that optimise this process of finding the scores.

Step II - applying IPA/DGM

The application of the IPA or the DGM to scores found in the first step is usually fast. For 11-node networks, the IPA took around 30 seconds using the software GOBNILP[3], and the DGM provided the graphical structure almost instantly in software R, on an Intel 2.83Ghz Core2 Quad CPU with 8GB RAM.

3.4 A Comparison with Some Other Methods

In this section, we compare the LMDM to other methods used to discover connectivity. Some methods are described in Section 2.3.

[1] `http://www.r-project.org/`
[2] `http://www.open-std.org/jtc1/sc22/wg14/`
[3] `http://www.cs.york.ac.uk/aig/sw/gobnilp/`

The number of nodes (n)	Sample Size (T)			
	100	200	600	1200
3	0.21	0.44	1.37	2.88
4	0.56	1.10	3.29	6.69
6	3.42	6.82	20.67	42.18
10	98.96	199.61	603.66	1099.72
11	167.99	325.50	1001.87	1982.08

Table 3.2: The time in minutes to find the scores for different numbers of nodes and sample sizes, and the discount factor was chosen in the range from 0.5 to 1.0 with increments of 0.01.

The Bayesian Network and The Dynamic Bayesian Network

The LMDM is a dynamic version of the Gaussian BN, where, unlike the latter, the LMDM allows the strength of connectivity to change over time. By explicitly modelling drift in the directed connection parameters, the LMDM can discriminate between models whose graphs are Markov equivalent. As a result two models, indistinguishable as BN models, become distinct when generalised into LMDMs, see examples of this in Chapter 4. The directed edges of the LMDM can be tentatively associated with a potential causal directionality, as argued in Queen and Albers (2009), and hint at the *effective* connectivity rather than the *functional* connectivity. The dynamic version of the BN is the DBN. This uses a Vector Autoregression (VAR) type time series sequentially rather than the state space employed in the LMDM, which can also be used to represent certain Granger causal hypotheses.

The Dynamic Granger Causality

Possibly the closest family of competitive models are the dynamic Granger causal models (Havlicek *et al.*, 2010). However, in general, the scores of these models are not factorable and so are much slower to search over. Moreover, comparing this with the simplest form of the LMDM we note that regression relationship are lagged whereas in the LMDM they are contemporaneous. Also, the DGC can be seen as the multivariate DLM (MVDLM) where the covariates are the past values of dependent variables. Queen *et al.* (2008) compared the LMDM with the MVDLM. They concluded that the forecast performance of the former was better, considering two model selection measures: the mean squared error (MSE) and the mean absolute deviation (MAD). Note that the LMDM is fitted based only on the individual

observational variances whilst in the MVDLM it is necessary to estimate the covariance matrix of time series variables and so the latter model is more complex.

Other dynamic models

In the previous neuroimaging literature, a sliding time window has been used to estimate the dynamic correlation among brain regions (Chang and Glover, 2010; Allen *et al.*, 2012; Leonardi *et al.*, 2013). Some methods also investigate the change points in network structures (*e.g.* Cribben *et al.*, 2012, considered sparse undirected graphs whilst Zhang *et al.*, 2013, considered the global structure, consisting of a global chain and V dependences among three networks). In contrast to LMDM, these methods study functional connectivity, and also use sophisticated but much more complex statistical computational algorithms rather than conditional conjugate analyses to perform inference.

The other methods, such as LDS, BDS and DCM that we reviewed in Section 2.3, are more sophisticated but also far too complicated to effectively score quickly enough over a large model space. Consequently these are not good candidates for use in the initial exploratory search we have in mind here. An important difference between these methods and LMDM is that while the dynamic of connectivity is directly estimated in LMDM, most other models consider connectivity as static or estimate only the different strengths of connectivity when modelling a different experimental situation. We show in our analyses below that in practice these strengths seem to drift in time.

Of course some authors discuss the possibility of a connectivity for each time point, $u_t = t$, including another dynamic system for the connectivity, *i.e.* $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{wa}_t$, where $\mathbf{wa}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Wa})$ and recall that $\mathbf{A}_t$ is the effective connectivity matrix in the LDS/BDS model (Bhattacharya *et al.*, 2006; Smith, J.F. *et al.*, 2011). But the inferential techniques needed for these models are considerably more complicated than for LMDM (*e.g.* using the Gibbs sampler scheme), and there are some extra assumptions they need to make, *e.g.* a fixed variance $\mathbf{Wa}$, which are not assumed in LMDM. Due the modeled drift in the connectivity, the LMDM is also able to detect change points in a straightforward manner and so to verify what changes in the interaction among brain regions over time. Such change points are beginning to be studied within this application (see Section 3.5.3). Moreover, unlike other models, the observational variances from the LMDM can be allowed to vary over time (see

Section 3.5.3).

Bhattacharya *et al.* (2006) proposed a similar model for LDS, but with the assumption that the observational errors are temporally dependent. This simpler class can also be implemented as a DLM (and so as an LMDM), as shown by West and Harrison (1997, Chapter 9). Another modification in the model assumption was suggested by Bhattacharya and Maitra (2011). These authors proposed the autoregressive model for effective connectivity as $\mathbf{A}_t = \rho\mathbf{A}_{t-1} + \mathbf{w}\mathbf{a}_t$, where ρ is a parameter to be estimated. So implicitly here we are using a random walk model for effective connectivity, *i.e.* $\mathbf{G}_t = \mathbf{I}_p$. It would be possible to use $\mathbf{G}_t = \rho\mathbf{I}_p$ and a similar procedure provided by Petris *et al.* (2009, Chapter 4) to estimate the matrix $\mathbf{G}$ within our method. However, this again gives rise to further complexities and certain computational issues.

The Time Varying Undirected Graph

Two main differences between the TVUG and the LMDM are firstly the former estimates the functional connectivity with undirected edges whilst the LMDM estimates the effective connectivity. Secondly, although the TVUG estimates the dynamic pattern of the connectivity, in contrast to the LMDM, it assumes that the time series variables are independent over time, and they follow the multivariate Gaussian distribution.

Other issues

One big advantage of the LMDM is that the estimation process is straightforward and closed conditional on setting the values of the discount factors. More explicitly its predictive distributions are products of Student t distributions in which the hyperparameters are found through well-known Kalman Filter algorithm. Thus, it is easy to find the posterior distribution of effective connectivity and so to test hypotheses about the parameters. Moreover, from the forecast distribution, the Bayes factor allows us to select a model directly. In contrast, other models need to use approximate inferential methods such as an Expectation maximization algorithm (EM) or Variational Bayes (VB) that are still quite difficult to implement for these classes and add other problems to the model selection process. They also often use a bootstrap analysis to verify if the effective connectivity is significant dramatically slowing down any search. Thus, searching over these classes becomes more difficult and time

consuming: a particular problem is that here we are selecting from a large set of alternative hypotheses. So the LMDM provides a very promising fast and informative explanatory data analysis of the nature of the dynamic network.

Another important advantage of the LMDM over most of its competitors is that it comes with a customized suite of diagnostic methods. Therefore, as in all modelling processes, a crucial step is to lead a diagnostic study and so to verify whether the results are valid. In the next section, we propose some diagnostic measures which are especially pertinent to model checking in this particular application. Violations detected through these diagnostics enable us to embellish the class to accommodate different features. For example, we are able to include time-dependent error variances, change points, interaction terms in the regression and so on within the models to better reflect the underlying model and refine the analysis. In Chapter 4, we demonstrate the usefulness of some of these methods to detect any deviations from the simplest LMDM and default and how to improve the model.

3.5 Diagnostic Analysis

In the past, Cowell *et al.* (1999) have convincingly argued that when fitting graphical models, it is extremely important to customize diagnostic methods, not only to determine whether the model appears to be capturing the data generating mechanism well but also to suggest embellishments of the class that might fit better. Their preferred methods are based on a one-step ahead prediction. We use these here. They give us a toolkit of sample methods for checking to see whether the best fitting model we have chosen through our selection methods is indeed broadly consistent with the data we have observed. We modified their statistics to give analogous diagnostics for use in our dynamic context. We give three types of diagnostic monitor, based on analogues for probabilistic networks (Cowell *et al.*, 1999).

First the *global monitor* is used to compare networks. After identifying a DAG providing the best explanation over the LMDM candidate models, the predicted relationship between a particular node and its parents can be explored through the *parent-child monitor*. Finally the *node monitor* diagnostic can indicate whether the selected model fits adequately. If this is not so, then a more complex model will be substituted, as illustrated below.

3.5.1 Global Monitor

The first stage of our analysis is to select the best candidate DAG using simple LMDMs, as described in Section 3.3. It is well known that the prior distributions on the hyperparameters of candidate models sharing the same features must first be matched (Heckerman, 1999). In this way, the BF techniques can be successfully applied in the selection of non-stochastic graphs in real data. If this is not done, then one model can be preferred to another, not for structural reasons but for spurious ones. This is also true for the dynamic class of models we fit here.

However, fortunately, the dynamic nature of the class of the MDM actually helps dilute the misleading effect of any such mismatch because, after a few time steps, evidence about the conditional variances and the predictive means is discounted, and the marginal likelihood of each model usually repositions itself, as we showed in Section 3.2.3. In particular, the different priors usually have only a small effect on the relative values of subsequent conditional marginal likelihoods. We describe below how we have nevertheless matched priors to minimise this small effect in the consequent Bayes factor scores driving the model selection.

Just as for BN to match priors, we can exploit a use decomposition of the Bayes factor score for the MDMs. By equation 3.6, the joint log predictive likelihood can be written as the sum of the log predictive likelihoods for each observation series given its parents: a *modularity* property (Heckerman, 1999). This assumption says that the predictive likelihood of a particular node depends only on the graphical structure, *i.e.*, $p(\mathbf{y}(i)|m_1) = p(\mathbf{y}(i)|m_2)$, if the set of parents of node i in m_1 is the same as in m_2. Therefore, when some features are incorporated within the model class, the *relative* scores of such models only discriminate the components of the model where they differ. Thus, again consider the graphical structure in Figure 3.2. For instance, suppose the LMDM is updated because node 3 exhibits heteroscedasticity. On observing this violation, the conditional one-step forecast distribution for node 3 can be replaced by one relating to a more complex model. The log Bayes factor comparing the original model with a heteroscedastic model is calculated as

$$\log\text{BF} = \sum_{t=1}^{T} \log p(y_t(3)|\mathbf{y}^{t-1}, y_t(1), y_t(2)) - \sum_{t=1}^{T} \log p^*(y_t(3)|\mathbf{y}^{t-1}, y_t(1), y_t(2)),$$

where $p^*(y_t(3)|\mathbf{y}^{t-1}, y_t(1), y_t(2))$ is the new one-step ahead forecast density for node 3 (see details about this distribution in Section 3.5.3). We set prior densities over the same component parameters over different models, because the model structure is common for all other nodes. The BF then discriminates between two models by finding the one that best fits the data *only* from the component where they differ: in our example the component associated with node 3. Even in larger scale models like the ones we illustrate below, we can therefore make a simple modification of scores in order to use the IP algorithm derived above, and in this way adapt the scores over graphs almost instantaneously.

In this setting we have found that the distributions for hyperparameters of different candidate parent sets is not critical for the BF model selection, provided that early predictive densities are comparable. We have found that a very simple way of achieving this is to set the prior covariance matrices over the regression parameters of each model a priori so that they are independent with a shared variance. Note the hyperparameters and the parameter δ of the nodes 1, 2 and 4 were the same for both models: homoscedastic and heteroscedastic for node 3. Many numerical checks have convinced us that the results of the model selection we describe above are insensitive to these settings *provided* that the high scoring models pass various diagnostic tests some of which we discuss below.

Other model selection measures may also have the modularity property. For instance, mean absolute deviation (MAD) and mean square error (MSE) are popular measures used in BN models (see *e.g.* Sun and Zhang, 2006; Zou and Feng, 2009; Le and Doctor, 2011). These measures are defined as

$$\begin{aligned} \text{MAD} &= \frac{1}{nT}\sum_{r=1}^{n}\sum_{t=1}^{T} \mid y_t(r) - f_t(r) \mid, \text{ and} \\ \text{MSE} &= \frac{1}{nT}\sum_{r=1}^{n}\sum_{t=1}^{T}(y_t(r) - f_t(r))^2, \end{aligned}$$

where $f_t(r) = \mathbb{E}\{Y_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t(r)\}$ (equation (3.5)). Coming back to the example of the problem of heteroscedasticity for node 3, it is notable that to compare MAD/MSE considering all nodes is the same as comparing them considering only the information for node 3, *i.e.*

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(y_t(3) - f_t(3))^2,$$

(or similar to MAD), because the elements $f_t(1)$, $f_t(2)$ and $f_t(4)$ are the same for original and heteroscedastic model for node 3.

However, MAD and MSE take into account only the forecast accuracy while the BF also consider the forecast precision, *i.e.* both the location and the spread of the predictive likelihood are used in its calculation. Thus, MAD/MSE is not able to account for all of the features in the prediction other than the mean. The BF score uses all aspects of the model (especially the variance), and this property is particularly useful (see Denison *et al.*, 2002).

In the MDM, as MAD and MSE are usually chosen when the primary purpose of modelling is prediction, these measures should be evaluated based on the marginal forecast mean, $\overline{f_t}(r) = \mathbb{E}\{Y_t(r)|\mathbf{y}^{t-1}\}$, rather than the conditional forecast mean $f_t(r)$ (see *e.g.* Queen *et al.*, 2008). However, because $\overline{f_t}(r)$ is calculated as a function of the forecast mean of its parents (see equation (7.14) in Appendix A), this *new version* of MAD and MSE does not have the modularity property. Thus, as node 3 is the parent of node 4, the value of $\overline{f_t}(4)$ depends on the marginal forecast mean of node 3 which may change from original to heteroscedastic model. The lack of this property will complicate both the search network and the diagnostic process. As the analysis of fMRI data focuses on identifying and estimating the connectivity, the BF is then more suitable for this application.

3.5.2 Parent-child Monitor

Because of modularity property, the relationship between a particular node and its parents can be assessed considering only this component in the MDM.

Let $Pa(r) = \{\mathbf{Y}_{pa(r)}(1), \ldots, \mathbf{Y}_{pa(r)}(p_r - 1)\}$, then

$$\log \mathrm{BF}_{ri} = \log p(\mathbf{y}(r)|Pa(r)) - \log p(\mathbf{y}(r)|Pa(r) \setminus \mathbf{y}_{pa(r)}(i)), \tag{3.11}$$

for $r = 1, \ldots, n$ and $i = 1, \ldots, p_r - 1$; where $\{Pa(r) \setminus \mathbf{Y}_{pa(r)}(i)\}$ means the set of all parents of $\mathbf{Y}(r)$ excluding the parent $\mathbf{Y}_{pa(r)}(i)$.

It is notable that this parent-child monitor can also be applied to verify the inclusion of a set of parents, $Pa^x(r)$ say, where $Pa^x(r) \subset \mathbf{X}_t(r)$, so that the new set of parents for node r is $Pa(r)^* = Pa^x(r) \bigcup Pa(r)$. Thus, the model selection criterion is now written as

$$\log \mathrm{BF}_{rx} = \log p(\mathbf{y}(r)|Pa(r)) - \log p(\mathbf{y}(r)|Pa^*(r)). \tag{3.12}$$

For instance, considering the graphical structure of Figure 3.2, equation (3.11) can be used to confirm the parent-child relationship between nodes 1 and 3, *i.e.*

$$\log \mathrm{BF}_{31} = \log p(\mathbf{y}(3)|\mathbf{y}(1), \mathbf{y}(2)) - \log p(\mathbf{y}(3)|\mathbf{y}(2)).$$

However, equation (3.12) can be used to assess whether node 1 is the parent of node 4, *i.e.*

$$\log \mathrm{BF}_{41} = \log p(\mathbf{y}(4)|\mathbf{y}(3)) - \log p(\mathbf{y}(4)|\mathbf{y}(3), \mathbf{y}(1)).$$

3.5.3 Node Monitor

Again the modularity ensures that the model for any given node can be embellished based on residual analysis. For instance, consider a non-linear structure for a root node r, *i.e.* one with no parents. On the basis of the partial autocorrelation of the residuals of the logarithm of the series, a more sophisticated model of the form

$$\log Y_t(r) = \theta_t^{(1)}(r) + \theta_t^{(2)}(r) \log Y_{t-1}(r) + v_t(r), \tag{3.13}$$

suggests itself. Note that this model still provides a closed form score for these components. The lower scores and their corresponding model estimation can then be substituted for the original steady models to provide a much better scoring dynamic model, but they still respect the same causal structure as in the original analysis.

Denoting $\log Y_t(r)$ by $Z_t(r)$, the conditional one-step forecast distribution for $Z_t(r)$ can then be calculated using a DLM on the transformed series $\{Z_t\}$. More generally if we hypothesize that $Z_t(r)$ can be written as a continuous and monotonic function of $Y_t(r)$, say $g(.)$, then the conditional one-step forecast cumulative distribution for $Y_r(r)$ can be found through

$$\begin{aligned} F_{Y_t(r)}(y) &= p(Y_t(r) \leqslant y|\mathbf{y}^{t-1}, Pa(r)) \\ &= p(Z_t(r) \leqslant g^{-1}(y)|\mathbf{y}^{t-1}, Pa(r)) \\ &= F_{Z_t(r)}(g^{-1}(y)). \end{aligned}$$

So $p^*(y_t(r)|\mathbf{y}^{t-1}, Pa(r))$, the conditional one-step forecast density for $Y_t(r)$ for this

new model can be calculated explicitly (see details in West and Harrison, 1997, section 10.6). It is also possible to keep the previous time series $\mathbf{Y}^t(r)$ as Gaussian and then simply regress on terms like $y^2 \cos y$, $y \sin y$ and so on for previous y. The predictive distribution of the models with these non-linear functional relationships still has a closed form in the DLM and so in the MDM.

This embellishment in the model for node r impacts on the estimation of effective connectivity between this node and its parents, but not for other connections. That is, the conditional forecast distribution and the posterior distribution of regression parameters for other nodes are independent of the model for node r, as they are calculated based on the observed values of variable $Y_t(r)$. However, the estimation of the marginal forecast parameters for a particular node depends on the forecast distribution of its parents. Therefore, any change in the distribution of $Y_t(r)$ may modify the marginal forecast distribution of the variables of its children. The idea here is that the better the prediction for a particular node, the better the prediction for its children, even though there is no intervention in the model of its children (Queen and Albers, 2009).

Recall that when the variance is unknown, the conditional forecast distribution is a noncentral t distribution with a location parameter $f_t(r)$, scale parameter $Q_t(r)$, and degrees of freedom, $n_{t-1}(r)$ (equation (3.5)). The one-step forecast errors are defined as $e_t(r) = Y_t(r) - f_t(r)$ and the standardized conditional one-step forecast errors as $se_t(r) = e_t(r)/Q_t(r)^{(1/2)}$. The assumption underlying the DLM is that the standardized conditional one-step forecast errors have an approximate Gaussian distribution, when $n_{t-1}(r)$ is large, and they are serially independent with constant variance (West and Harrison, 1997; Durbin and Koopman, 2001). These assumptions can be checked by looking at some graphs, such as QQ-plot, standardized residuals versus time, cumulative standardized residuals versus time and ACF-plot (Smith, 1985; Harrison and West, 1991; Durbin and Koopman, 2001). How to detect and to solve some problems are shown below.

Normality

The assumption of the standardized residuals have a Gaussian distribution can be checked by the quantile-quantile plot (*QQ-plot*). This graph plots the ordered standardized errors versus their theoretical quantiles, and so the closer the residual plots to a straight line,

the greater the evidence for normality (Durbin and Koopman, 2001).

Non-normality can also be associated with apparent changes in the observational variance (West and Harrison, 1997, Section 10.6; Anacleto Junior *et al.*, 2013a). For instance, the variance of a Poisson distribution equals the mean, and so the observational variance varies over time. The usual solution is a data transformation, *e.g.* the logarithm function, and then the original distribution becomes symmetry and closer to Gaussian distribution. However sometimes it is not easy to find a suitable transformation, and moreover, the model loses in terms of interpretability. In this case, a possible solution is to use a variance law (see below in heteroscedasticity).

Non-linear relationship between parent and child

The assumption of a linear relationship between parent and child can be checked by a scatterplot between the observation values of their variables. For instance, suppose Figure 3.8 provides the relation between node 1 and its parent, node 2. There is no a *global* linearity between $\mathbf{Y}(1)$ and $\mathbf{Y}(2)$, and the connectivity strength (also the intercept) changes according to the values of variables (*e.g.* the slope parameter of the blue line is smaller than of the green line). But, as the DLM can deal with changes in the regression parameters, this class of model assumes *local* linearity rather than global linear relation (West and Harrison, 1997). However, in practice, it is expected that the parameters change slowly and smoothly over time and, in this way, nonlinear regression problems should be solved. A possible solution is to modify the nonlinear relation so that a linear regression can be used. For instance, returning to the example in Figure 3.8. The relation between node 1 and node 2 may be written as:

$$Y_t(1) = \gamma_t \exp^{\beta_t Y_t(2)} \exp^{v_t(r)} .$$

Thus, in order to fit data by linear MDM, the logarithm function can be used as such

$$\log Y_t(1) = \theta_t^{(1)}(1) + \theta_t^{(2)}(1) Y_t(2) + v_t(1),$$

where $\theta_t^{(1)}(1) = \log \gamma_t$ and $\theta_t^{(2)}(1) = \beta_t$. As other example, Santos (2014) proposed a class of models, the Gaussian Dynamic Bayesian Smooth Transition Autoregressive (DBSTAR),

which analyses nonlinear autoregressive time series processes using the autoregressive formulations of DLMs.

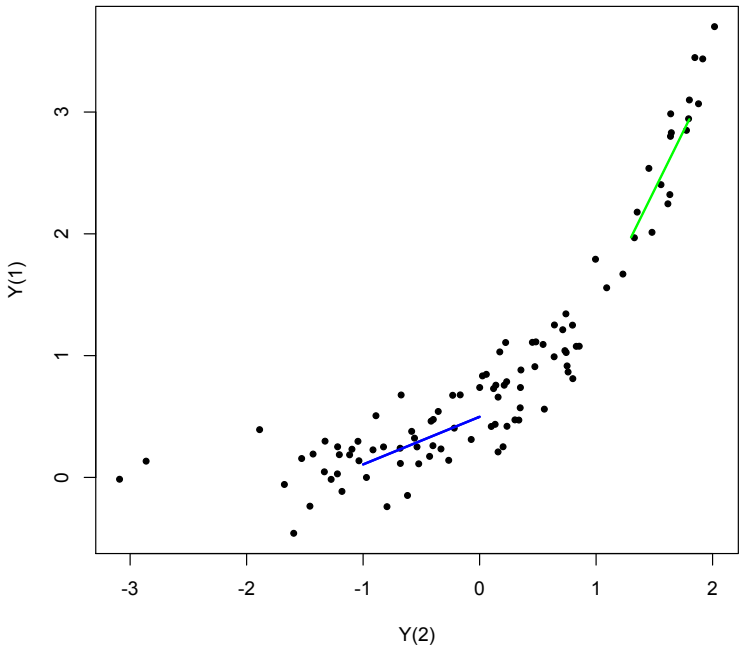


Figure 3.8: The local linear relation between node 1 and its parent, node 2.

Heterocedasticity

Heteroscedasticity can be checked visually through the graph of standard residuals against the time (Smith, 1985), *e.g.* an hourglass shape indicates the variance is not constant. Or yet the homoscedasticity assumption may not be verified when the boxplots of exclusive subsets of a particular observational variable show different patterns (Anacleto Junior *et al.*, 2013a).

In order to address this problem, a random variance, say $k_{tr}(.)$, can be incorporated into the model so that the observational variance at time t, $V_t(r)$, is expected to be the product between $k_{tr}(\mathbf{F}_t(r)'\boldsymbol{\theta}_t(r))$ and $V(r)$. $k_{tr}(.)$ should be chosen according to data (Migon *et al.*, 2005), *e.g.* Anacleto Junior *et al.* (2013a) use $k_{tr} = \exp\{\beta \log(\mathbf{F}_t(r)'\boldsymbol{\theta}_t(r))\}$, for a specified β. In this way, the conditional one-step forecast distribution for node r can be shown to be $(Y_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t) \sim \mathcal{T}_{n_{t-1}(r)}(f_t(r), Q_t^h(r))$, where the parameters $f_t(r)$ and $n_{t-1}(r)$ are defined as equation (3.5), but $Q_t^h(r)$ is now defined as function of k_{tr} (West and Harrison,

1997, Section 10.7), as follows

$$Q_t^h(r) = k_{tr} S_{t-1}(r) + \mathbf{F}_t'(r)\mathbf{R}_t(r)\mathbf{F}_t(r).$$

Serial Correlation

To verify the assumption that the standardized conditional one-step forecast errors are serially independent, we can use *the autocorrelation function* and *the partial autocorrelation function*. The autocorrelation function (ACF) at lag k is defined as the Pearson correlation between $se_t(r)$ and $se_{t+k}(r)$, for $k = 0, \ldots, K$, where K should be smaller than $T/4$ (Box *et al.*, 1994, Section 3.2). The partial autocorrelation function (PACF) at lag k is the last regression coefficient in an autoregressive model of order k, say $\Phi_{kk}(r)$ in

$$se_t(r) = \Phi_{k1}(r)se_{t-1}(r) + \Phi_{k2}(r)se_{t-2}(r) + \ldots + \Phi_{kk}(r)se_{t-k}(r) + \varepsilon_t(r),$$

where $\varepsilon_t(r)$ is a white noise (Box *et al.*, 1994, Section 3.2). The patterns of the ACF and PACF graphs indicate the components of the autoregressive (AR) process and the moving average (MA) process that should be used in the model. Briefly, when (Box *et al.*, 1994, Section 6.2)

- ACF decays exponentially or displays some nonzero elements whilst the first elements of PACF are nonzero, then the order of autoregressive model is the last nonzero element of PACF;
- the first elements of ACF are nonzero whilst PACF decays exponentially, then the order of moving average model is the last nonzero element of ACF;
- both ACF and PACF decay exponentially, then autoregressive and moving average should be used;
- all elements of ACF and PACF are non-zero, then the assumption of serial independence can be considered.

The AR(p) components can be included in the DLM, supposing a root node r, as

follows

$$Y_t(r) = \theta_t^{(1)}(r) + \theta_t^{(2)}(r)Y_{t-1}(r) + \ldots + \theta_t^{(p+1)}(r)Y_{t-p}(r) + v_t(r).$$

Of course, if it is necessary, the set of parents of node r can also be included in the observation equation above. Note that this still provides distributions in a closed form.

West and Harrison (1997, Chapter 9) provided one way to include the ARMA(p,q) components in DLM form. The usual ARMA observation equation is

$$Y_t(r) = \sum_{i=1}^{p} \Phi_{ti}(r)Y_{t-i}(r) + \sum_{j=1}^{q} \psi_{tj}(r)v_{t-j}(r) + v_t(r).$$

This can be rewritten in the form:

$$Y_t(r) = \mathbf{F}_t(r)'\boldsymbol{\theta}_t(r),$$

where $\mathbf{F}_t(r) = (1, 0, \ldots, 0)$ and the first element of the $u-$vector $\boldsymbol{\theta}_t(r)$ is $Y_t(r)$, u is defined as $max(p, q+1)$, so that $\Phi_{ti}(r) = 0$ for $i > p$ and $\psi_{tj}(r) = 0$ for $j > q$. Note that there is no white noise in this observation equation. The state equation is written in original form, *i.e.*

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t,$$

but considering

$$\mathbf{G}_t = \begin{bmatrix} \Phi_{t1}(r) & 1 & 0 & \ldots & 0 \\ \Phi_{t2}(r) & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Phi_{tu-1}(r) & 0 & 0 & \ldots & 1 \\ \Phi_{tu}(r) & 0 & 0 & \ldots & 0 \end{bmatrix}$$

and $\mathbf{w}_t(r)' = (1, \psi_{t1}(r), \ldots, \psi_{tu-1}(r))v_t(r)$. The innovation variance matrix is defined as $\mathbf{W}_t(r) = W(r)(1, \psi_{t1}(r), \ldots, \psi_{tu-1}(r))'(1, \psi_{t1}(r), \ldots, \psi_{tu-1}(r))$. Finally a random walk model can be specified for $\mathbf{G}_t$, so that $\mathbb{E}\{\mathbf{G}_t|\mathbf{G}_{t-1}, \boldsymbol{\theta}_{t-1}, \mathbf{y}^{t-1}\} = \mathbf{G}_{t-1}$. More details about the ARMA components in DLM can be seen in West and Harrison (1997, Chapter 9). In addition, if data exhibit cycles, it is possible to deal with them including cubic splines into the LMDM, as suggested by Anacleto *et al.* (2013b).

Influence measures

The influence of individual observations on a model analysis can be verified comparing the posterior distribution $p(\boldsymbol{\theta}_t(r), \phi(r)|\mathbf{y}^T(r))$ with *the jackknifed posterior distribution* $p(\boldsymbol{\theta}_t(r), \phi(r)|\mathbf{y}^{T\backslash t}(r))$, where $\mathbf{y}^{T\backslash t}$ is the information set without the observation $\mathbf{y}_t$, *i.e.* $\mathbf{y}^{T\backslash t} = (\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{y}_{t+1}, \ldots, \mathbf{y}_T)$, for $t = 1, \ldots, T$. Therefore, Harrison and West (1991) used the following Kullback-Leibler divergence measure to compare these two distributions

$$K_t(r) = \int\int \log\left\{\frac{p(\boldsymbol{\theta}_t(r), \phi(r)|\mathbf{y}^{T\backslash t})}{p(\boldsymbol{\theta}_t(r), \phi(r)|\mathbf{y}^T)}\right\} p(\boldsymbol{\theta}_t(r), \phi(r)|\mathbf{y}^{T\backslash t}) d\boldsymbol{\theta}_t(r) d\phi(r).$$

Using the smoothed posterior distribution (see Section 3.2.2), *i.e.*

$$\begin{aligned}(\boldsymbol{\theta}_t(r)|\mathbf{y}^T) &\sim \mathcal{T}_{n_T(r)}\left(\mathbf{sm}_t(r), \mathbf{sC}_t(r)\right), \\ (\phi(r)|\mathbf{y}^T) &\sim \mathcal{G}\left(\frac{n_T(r)}{2}, \frac{d_T(r)}{2}\right),\end{aligned}$$

Harrison and West (1991) provided computationally simple equations in order to find the values of $K_t(r)$, as follows.

$$\begin{aligned}
K_t(r) &= I_t(r) + J_t(r), \\
2I_t(r) &= \omega_t(r) - 1 - \log\omega_t(r) + (\omega_t(r) - 1)U_t(r), \\
2J_t(r) &= U_t(r) - n_T(r)\log\left(1 + \frac{U_t(r)}{n_T(r) - 1}\right) - \gamma\left(\frac{1}{2}n_T(r) - \frac{1}{2}\right) + 2\log\left(\frac{\Gamma\left(\frac{1}{2}n_T(r)\right)}{\Gamma\left(\frac{1}{2}n_T(r) - \frac{1}{2}\right)}\right), \\
\omega_t(r) &= s_t^*(r)/q_t(r), \\
s_t^*(r) &= S_T(r)\left(\frac{n_T(r) - (y_t(r) - g_t(r))^2/q_t(r)}{n_T(r) - 1}\right), \\
q_t(r) &= S_T(r) - \mathbf{F}_t'(r)\mathbf{sC}_t(r)\mathbf{F}_t(r), \text{ also } q_t(r) > 0, \\
g_t(r) &= \mathbf{F}_t'(r)\mathbf{sm}_t(r), \\
U_t(r) &= e_t^{*2}(r)/\left(s_t^*(r)\omega_t(r)\right), \\
e_t^*(r) &= y_t(r) - g_t^*(r), \\
g_t^*(r) &= g_t(r) + (\omega_t(r) - 1)\left(g_t(r) - y_t(r)\right),
\end{aligned}$$

$\gamma(.)$ is the digamma function, *i.e.* the logarithmic derivative of the gamma function. The component $I_t(r)$ measures the differences between the posterior distributions for $\boldsymbol{\theta}_t(r)$ us-

ing the full and the remaining data, while $J_t(r)$ measures the difference between the two inverse gamma posteriors for $V(r)$. The standardized jackknifed residual $U_t(r)$ measures the distance between $y_t(r)$ and predictions based on the remaining data. On the other hand, the element $\omega_t(r)$ measures the relative distance between $\mathbf{F}_t$ and the remaining regression vectors (Harrison and West, 1991).

Before deciding what should be done to the observations that have high influence on model fit, we should be aware of the causes that lead to their appearance. In many cases the reason for their existence can determine how to address these observations. The simplest way is to eliminate them. In this case, the posterior distribution at time t is the same as the prior, *i.e.* $(\boldsymbol{\theta}_t(r)|\mathbf{y}^t(r)) \sim \mathcal{T}_{n_t(r)}(\mathbf{m}_t(r), \mathbf{C}_t(r))$, where $\mathbf{m}_t(r) = \mathbf{m}_{t-1}(r)$, $\mathbf{C}_t(r) = \mathbf{R}_t(r)$ and $n_t(r) = n_{t-1}(r)$. However it is important to treat carefully these observations as they may contain important information about the underlying data.

We will show in the next item, change points, that there is another way to investigate the influence of individual observations in dynamic models, using intervention through the forecast and the filtered distributions. In contrast, here we assessed the agreement between a model and a particular individual observation, considering the entire time series, *i.e.* the smoothed distributions. This method was developed specifically for the class of DLM, but based on the same idea used in diagnostic analysis for linear models (see *e.g* Smith and Pettit, 1985, and Bruce and Martin, 1989). That is, this method separates out the effect of a particular observation from the inferences for the regression parameters, considering all other observations (Harrison and West, 1991). Therefore, we compare the distribution used to make inferences for the model parameters, considering all observations (or, in the context of dynamic models, all times), with the distribution free the influence of $Y_t(r)$. Many divergence measures can be used to make this comparison (see *e.g.* Johnson and Geisser, 1983, Bernardo, 1985, and Smith and Pettit, 1985). The divergence measure proposed by Harrison and West (1991), and shown here, uses in this calculation the moments of the two distributions, smoothed and jackknifed, and as it has a closed form, it is computationally easy to find the results. Moreover, its calculation is informative, because its components $I_t(r)$ and $J_t(r)$ can be interpreted as mentioned above.

Change points

For some reason, the time series model may not fit the data well from a particular time point, and then the intervention may be necessary. Interventions can be classified as *feed-forward* or *feed-back* (West and Harrison, 1997, Chapter 11). The former is based on external information, so that the model is changed at time t' in order to prevent estimation or forecast problems for $t > t'$. Moreover, the intervention precedes the observation of the series at time t'. For instance, in traffic flow networks, if a particular road is closed, then there may be a sharp decrease in the traffic of its child road, and so this information can be included in the forecast model (Queen and Albers, 2009). In the context of task design fMRI experiment, for example, we can set up informative priors as known change points. In contrast, the feed-back intervention is applied when a monitoring process detects deteriorations in forecasting performance. This can be seen as corrective action, and the monitoring is basically based on the model selection measures, see below.

The change in the regression parameters, which may lead to poor predictive performance, may have several causes. In fMRI studies, change points may be explained by psychological processes, such as stress or anxiety during the experiment (see *e.g.* Robinson *et al.*, 2010; Aston and Kirch, 2012). West and Harrison (1997) recommended that interventions should be made when forecasting problems are detected even though the causes are not identified. Therefore, a way to deal with these problems is to increase uncertainty at the time point specified by the monitoring, and then the model is updated to adjust to these new data (West and Harrison, 1997). That is, the system equation with intervention is written as

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{w}_t + \boldsymbol{\xi}_t,$$

where $\mathbf{w}_t$ is defined as before, *i.e.* $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t)$ and $\mathbf{W}_t = \text{blockdiag}\{\mathbf{W}_t(1), \ldots, \mathbf{W}_t(n)\}$; $\boldsymbol{\xi}_t$ is an error that represents an intervention in the parameter distributions and defined as $\boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{h}_t, \mathbf{H}_t)$, where $\mathbf{h}_t$ represents the expected change in the regression parameters whilst $\mathbf{H}_t$ represents the uncertainty about these changes. Formally $\mathbf{h}'_t = (\mathbf{h}_t(1)', \ldots, \mathbf{h}_t(n)')$ and $\mathbf{h}_t(r)$ is p_r-dimensional vector of node r at time t. $\mathbf{h}_t(r)$ can be defined as a function of the expected value of $\boldsymbol{\theta}_t(r)$ with intervention, such as $\mathbf{h}_t(r) = \mathbb{E}\left[\boldsymbol{\theta}_t(r)|\mathbf{y}^{t-1}, \text{intervention}\right] - \mathbf{m}_{t-1}(r)$. The intervention variance is defined as $\mathbf{H}_t = \text{blockdiag}\{\mathbf{H}_t(1), \ldots, \mathbf{H}_t(n)\}$ and each $\mathbf{H}_t(r)$ being a $p_r \times p_r$ matrix. Note that some parameters may be expected not to change,

and so their respective elements in $\mathbf{h}_t(r)$ and in the diagonal (and the corresponding elements off-diagonal) of matrix $\mathbf{H}_t(r)$ can be zero to protect these parameters from intervention.

Usually $\mathbf{h}_t(r)$ is defined as zero vector whilst $\mathbf{H}_t(r)$ has large values. Thus, it allows the model to adapt better to these data from a different system. The model is self-corrected because as the prior variances of regression parameters increase at time when change was detected, then further data have more influence in the posterior distributions. Consider the reparameterisation $\mathbf{H}_t(r) = V(r)\mathbf{H}^*_t(r)$ and $\mathbf{W}^*_{ht}(r) = \mathbf{W}^*_t(r)+\mathbf{H}^*_t(r)$. In practice, as $\mathbf{W}^*_{ht}(r)$ is expected to be large when there is intervention at time t, then it can be represented by a small discount factor. Thus, the filtering algorithm can be written as

$$\mathbf{W}^*_{ht}(r) = \frac{1-\delta^*_t(r)}{\delta^*_t(r)}\mathbf{C}^*_{t-1}(r),$$

where $\delta^*_t(r) = \delta(r)$, the usual discount factor, if there is no intervention at time t, or $\delta^*_t(r)$ is a small value for intervention, *e.g.* 0.1.

The model selection criteria, Bayes factor, can be used to verify changes in the model fitting (West and Harrison, 1997). Defining the current model as M_0 and an alternative model as M_1, the logarithm of BF at time t for node r is defined as

$$\begin{aligned}\log \mathrm{BF}_t(r) &= \log\left[\frac{p(y_t(r)|\mathbf{y}^{t-1},\mathbf{x}_t(r),M_0)}{p(y_t(r)|\mathbf{y}^{t-1},\mathbf{x}_t(r),M_1)}\right]\\ &= \mathrm{LPL}_{rt}(M_0) - \mathrm{LPL}_{rt}(M_1).\end{aligned}$$

The cumulative logBF considering the last k observations is calculated as

$$\begin{aligned}\log \mathrm{BF}^k_t(r) &= \log\left[\frac{p(y_{t-k+1}(r),\ldots,y_{t-1}(r),y_t(r)|\mathbf{y}^{t-k},\mathbf{x}_t(r),M_0)}{p(y_{t-k+1}(r),\ldots,y_{t-1}(r),y_t(r)|\mathbf{y}^{t-k},\mathbf{x}_t(r),M_1)}\right]\\ &= \log\left[\frac{\prod_{i=t-k+1}^{t} p(y_i(r)|\mathbf{y}^{i-1},\mathbf{x}_i(r),M_0)}{\prod_{i=t-k+1}^{t} p(y_i(r)|\mathbf{y}^{i-1},\mathbf{x}_i(r),M_1)}\right]\\ &= \sum_{i=t-k+1}^{t}\log \mathrm{BF}_i(r).\end{aligned}$$

Define $L_t(r) = \min\limits_{1\le k\le t}\{\log \mathrm{BF}^k_t(r)\}$, and $l_t(r) = k$ so that $L_t(r)$ is the same as $\log \mathrm{BF}^k_t(r)$. As $\log \mathrm{BF}^k_t(r)$ measures the evidence for M_0 considering the last k observations, $L_t(r)$ reflects the value of the lowest possible evidence for the current model for node r whilst $l_t(r)$ indicates the period that provides this value. When $L_t(r)$ is much smaller than zero, then

there is strong evidence against the current model for node r. A predefined threshold τ is used in the monitoring process so that $L_t(r) < \tau$ indicates the intervention is necessary at time t. West and Harrison (1997) suggested that τ should be between -2.3 and -1.6.

Note that $\log \mathrm{BF}_t^1(r) = \log \mathrm{BF}_t(r)$ and $\log \mathrm{BF}_t^k(r) = \log \mathrm{BF}_t(r) + \log \mathrm{BF}_{t-1}^{k-1}(r)$ since

$$\begin{aligned} \log \mathrm{BF}_{t-1}^{k-1}(r) &= \sum_{i=t-k+1}^{t-1} \log \mathrm{BF}_i(r) \\ \log \mathrm{BF}_t(r) + \log \mathrm{BF}_{t-1}^{k-1}(r) &= \sum_{i=t-k+1}^{t} \log \mathrm{BF}_i(r). \end{aligned}$$

Therefore $L_t(r)$ is easily updated over time as

$$\begin{aligned} L_t(r) &= \min\left\{\log \mathrm{BF}_t(r), \min_{2\le k\le t} \log \mathrm{BF}_t^k(r)\right\} \\ &= \min\left\{\log \mathrm{BF}_t(r), \min_{2\le k\le t} \left[\log \mathrm{BF}_t(r) + \log \mathrm{BF}_{t-1}^{k-1}(r)\right]\right\} \\ &= \log \mathrm{BF}_t(r) + \min\left\{0, \min_{1\le j\le t-1} \log \mathrm{BF}_{t-1}^{j}(r)\right\} \\ &= \log \mathrm{BF}_t(r) + \min\left\{0, L_{t-1}(r)\right\}, \quad \text{for } t > 1. \end{aligned}$$

Moreover $l_t(r) = 1$ if $L_{t-1}(r) \ge 0$ or $l_t(r) = l_{t-1}(r) + 1$ if $L_{t-1}(r) < 0$.

When the plot of cumulative standardized residuals versus time does not show a random pattern, this may indicate that there are change points. The following monitoring algorithm can then be applied, beginning $t = 0$ (West and Harrison, 1997):

1. Perform the usual analysis, finding for example the forecast distribution for node r, $\log \mathrm{BF}_t(r)$, $L_t(r)$ and $l_t(r)$ for $t = t + 1$; and go to item 2.

2. If $\log \mathrm{BF}_t(r) < \tau$ or $L_t(r) < \tau$ or $l_t(r) > 3$ then there is evidence for model breakdown and go to item 3, otherwise go to item 1. You can decrease the sensitivity of this monitoring, excluding the condition $l_t(r) > 3$ or increasing the threshold for $l_t(r)$, *e.g.* $l_t(r) > 4$, or even dropping the value of τ.

3. Call the intervention, updating the Kalman-filter algorithm, using now $\delta_t^*(r) = 0.1$. Reinitiate the monitoring process, considering both $L_t(r)$ and $l_t(r)$ equal to zero. Go to item 1.

When $L_t(r) < \tau$ and $l_t(r) = 1$, then $Y_t(r)$ can be seen as a potential outlier or the

system began to change at time t. In contrast, if $L_t(r) < \tau$ and $l_t(r) > 1$, then possibly the change began small at $l_t(r)$ past time, and so the evidence against the current model increased over time. Note that, although this technique aims to improve the model fitting, it should be used carefully since the innovation variance increases at the points where the change was detected, and hence increasing the uncertainty of the smoothed estimates over entire time, as shown in Section 4.4.

This monitoring algorithm can be extended, considering 2 or more alternative forecast distributions. In this case, $\log \mathrm{BF}_{i,t}(r)$, $L_{i,t}(r)$ or $l_{i,t}(r)$ are calculated for every alternative model M_i and the intervention is done if the condition specified in the item 2 of monitoring algorithm is true for any $i = 1, \ldots, \mathcal{M}$, where $\mathcal{M}$ is the number of alternative models. These alternative models should be specified so that changes in the behaviour of time series can be detected. In the next chapter, we apply this theory into real datasets, considering that there is no connectivity as the alternative model.

This naive approach seems to deal adequately with identified change points, however other alternative distributions can be tested. For instance, we can use the fact that the assessment of the forecast distribution of $Y_t(r)$ is equivalent to the assessment of the standardized conditional one-step forecast errors, $se_t(r)$, as the latter is a linear function of the former. Therefore, the current model M_0 can be defined in terms of the standardized forecast distribution, *i.e.* $(se_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t, M_0) \sim \mathcal{N}(0, 1)$ whilst the alternative distribution M_1 can be defined as $(se_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t, M_1) \sim \mathcal{N}(h, 1)$ for a predefined value h. Thus it is not difficult to see that the $\log \mathrm{BF}_t(r) = 0.5(h^2 - 2hse_t(r))$ (West and Harrison, 1997).

Queen and Albers (2009) also suggested an intervention distribution including one more error term in the observation equation, as shown here for the system equation. In this case, the expected change occurs in the distribution of observations rather than the distribution of connectivity. As Queen and Albers (2009) were interested in making predictions of traffic flows, they estimated the marginal forecast parameters (rather than the conditional forecast distribution that we are using here), and so discussed the effects of intervention considered for one node into the marginal forecast of other nodes.

Chapter 4

The Evaluation Methodology

As we asserted in previous chapters, it is possible for an MDM to distinguish different directions of relationships in DAGs that are Markov equivalent in static analysis. Queen and Albers (2009) argued this idea using an intervention method and real datasets, and then reported that the directed edges of the MDM can be tentatively associated with a potential causal directionality.

In this chapter, we investigate the potential of an MDM to discriminate models with the same dependence constraints. In addition, we show that the MDM can distinguish between effective connectivities that change over time and those that remain the same, *i.e.* whether they originated from a dynamic or static process. We will explore these questions below and demonstrate how this is possible using a simulation experiment, in Section 4.1. In Section 4.2, we compare the performance of the MDM-IPA with other methods used to estimate connectivity, using the DCM fMRI and the MDM synthetic data.

In addition, we apply the search network methods and diagnostic measures, described above, using real fMRI data. Firstly, in Section 4.3, we analyse data which was obtained through a multivariate method, ICA. In the next section, our methods are applied to the data that have not been pre-processed, such as using the ICA. We then demonstrate the promise of the MDM for modelling typical fMRI datasets obtained in a variety of ways. Notice that there are a small number of nodes (three or four regions) in the datasets used here, but higher dimensional data are used in the next chapter, considering group analysis techniques.

4.1 The MDM Assessment

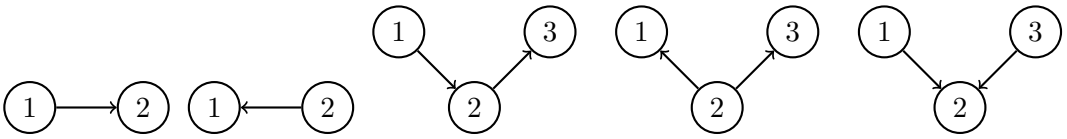
Here we simulated 100 datasets from every known MDM using sample sizes $T = 100$, 200 and 300, and different dynamic levels $\mathbf{W}^*(r) = 0\mathbf{I}_{p_r}$ (static), $0.001\mathbf{I}_{p_r}$, $0.01\mathbf{I}_{p_r}$ and $0.1\mathbf{I}_{p_r}$. The impact of these different scenarios on the MDM results was verified using 2 and 3 regions and each T and $\mathbf{W}^*(r)$ pair.

For two nodes, data were generated using the MDM with graph DAG1 given in Figure 4.1(a). The initial values for the regression parameters were 0.3 for connection between $Y(1)$ and $Y(2)$, *i.e.* $\theta_0^{(2)}(2)$, and the value 0 for other θ's (intercept parameters). The observational variance was defined as 12.5 for $Y(1)$ and 6.3 for $Y(2)$ so that the marginal variances were almost the same for both regions. Thus we set

$$\theta_{ti}^{(k)}(r) = \theta_{t-1i}^{(k)}(r) + w_{ti}^{(k)}(r), \quad w_{ti}^{(k)}(r) \sim \mathcal{N}(0, W^{(k)}(r)),$$

for $r = 1, \ldots, n$; $n = 2$; $t = 1, \ldots, T$; $i = 1, \ldots, 100$ replications; $k = 1, \ldots, p_r$; $p_1 = 1$; $p_2 = 2$; $W^{(k)}(r) = W^{*(k)}(r) \times V(r)$ and $W^{*(k)}(r)$ is the k^{th} element of the diagonal of matrix $\mathbf{W}^*(r)$ defined above. Observed values were then simulated using the following equations:

$$\begin{aligned} Y_{ti}(1) &= \theta_{ti}^{(1)}(1) + v_{ti}(1), \qquad v_{ti}(1) \sim \mathcal{N}(0, V(1)); \\ Y_{ti}(2) &= \theta_{ti}^{(1)}(2) + \theta_{ti}^{(2)}(2) Y_{ti}(1) + v_{ti}(2), \qquad v_{ti}(2) \sim \mathcal{N}(0, V(2)). \end{aligned}$$



(a) DAG1 (b) DAG2 (c) DAG3 (d) DAG4 (e) DAG5

Figure 4.1: Directed acyclic graphs used in the first synthetic study. With 2 nodes, (a) DAG1 and (b) DAG2 are Markov equivalent as static BNs. For 3 nodes, (c) DAG3 and (d) DAG4 are considered Markov equivalent whilst neither is equivalent to (e) DAG5.

For three nodes, the graphical structure used to obtain the synthetic data is shown in Figure 4.1(c), DAG3. The initial values for the regression parameters were 0.3 for the connectivity between $Y(1)$ and $Y(2)$, *i.e.* $\theta_0^{(2)}(2)$, 0.2 for the connectivity between $Y(2)$ and $Y(3)$, *i.e.* $\theta_0^{(2)}(3)$, and the value 0 for other θ's (intercept parameters). The observational variance was set to be the same as for two nodes for the first and second variables and 5.0 for

$Y(3)$. The observation and system equations were also set to be the same when considering 2 nodes, except for setting $n = 3$, $p_3 = 2$ and

$$Y_{ti}(3) = \theta_{ti}^{(1)}(3) + \theta_{ti}^{(2)}(3)Y_{ti}(2) + v_{ti}(3), \qquad v_{ti}(3) \sim \mathcal{N}(0, V(3)).$$

The log predictive likelihood (LPL) was first computed for different values of discount factor (DF or δ), using a weakly informative prior with $n_0(r) = d_0(r) = 0.001$ and $\mathbf{C}_0^*(r) = 3\mathbf{I}_{p_r}$ for all r. The discount factors were chosen as the value that maximized the LPL.

Dynamic X static models

To better understand the effect of estimating the connectivity when applying a wrong static/dynamic model, the largest static datasets ($\mathbf{W}^*(r) = 0\mathbf{I}_{p_r}$ and $T = 300$) were fitted with a dynamic model, using graph DAG1 and $\delta = 0.93$ (the average of DF in fitted models of the data generated with $\mathbf{W}^*(r) = 0.001\mathbf{I}_{p_r}$). Then, dynamic datasets from the same scenario, *i.e.* DAG1, $T = 300$, and $\mathbf{W}^*(r) = 0.001\mathbf{I}_{p_r}$, were fitted using the static model with $\delta = 1$. Figure 4.2 shows the true (blue lines) and smoothed estimated values of parameter $\theta_t^{(2)}(2)$ - connectivity $1 \to 2$, versus time t, for dynamic (violet lines) and static (green lines) models.

It can be seen that when the data are generated from a static model, dynamic models usually estimate the true values quite well, see Figure 4.2(a). Most of the times, they simply alternate between under and over-estimating the true values of parameters but nevertheless centre around the true value. In contrast, when the data are simulated from the dynamic models, as might be expected, static models fail to appropriately describe the series at each time point, as can be seen in Figure 4.2(b). This phenomenon is particularly pertinent to this application. Because we know connectivities change over time, by fitting static models (as we typically do when fitting BNs) we fit models that can score very poorly even when the topology of the connectivity is right! So in particular any preliminary model search using static BN models is likely to be unreliable and potentially misleading in this dynamically evolving environmental.

In order to verify whether the Bayes factor could distinguish between static and dynamic system structure, we plotted the histograms of the logBF that compared the Dynamic model ($\delta = 0.93$) with Static model ($\delta = 1$) for DAG1. We used 100 replications of a sample generated considering $T = 300$, and dynamic ($\mathbf{W}^*(r) = 0.001\mathbf{I}_{p_r}$) and static ($\mathbf{W}^*(r) = 0\mathbf{I}_{p_r}$)

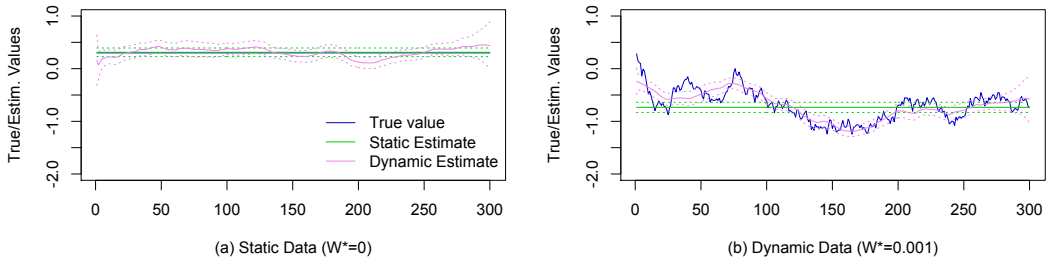


Figure 4.2: The true value (blue lines) and estimation result by smoothing for parameter $\theta_t^{(2)}(2)$ - connectivity $1 \rightarrow 2$, from DAG1, considering dynamic model (mean δ of 0.93 and violet lines) and static model (mean δ of 1 and green lines), for a particular replication. The dashed lines represent the 95% HPD intervals. (a) shows results from data simulated based on static model ($\mathbf{W}^*(r) = 0\mathbf{I}_{p_r}$) while (b) shows results for data from dynamic model ($\mathbf{W}^*(r) = 0.001\mathbf{I}_{p_r}$).

systems (see the left and the right histograms of Figure 4.3). The logBF shows strong evidence for the correct model in all situations, always being far from ± 1 in the appropriate direction.

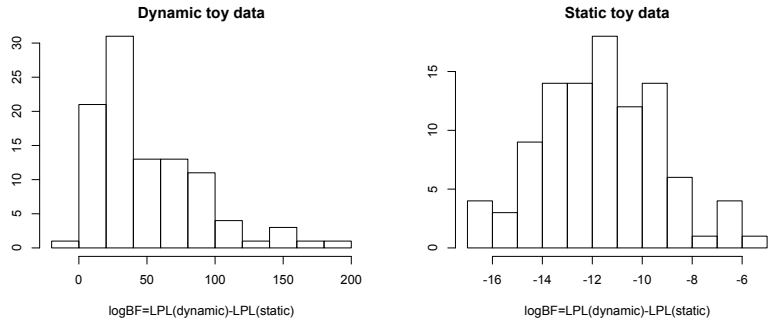


Figure 4.3: The histogram of logBF comparing dynamic model ($\delta = 0.93$) with static model ($\delta = 1$) for DAG1, considering 100 replications of a sample generated considering $T = 300$, via (left) dynamic ($\mathbf{W}^*(r) = 0.001\mathbf{I}_{p_r}$) and (right) static ($\mathbf{W}^*(r) = 0\mathbf{I}_{p_r}$) systems.

Figure 4.4 shows the log predictive likelihood versus different values of the discount factor, considering DAG3 (solid lines), DAG4 (dashed lines) and DAG5 (dotted lines). The sample size increases from the first to the last row whilst the dynamic level (innovation variance) increases from the first to the last column. Although the ranges of LPL differ across the graphs, the range sizes are the same, *i.e.* 500 so that it is easy to compare them. We can see in this figure that the choice of δ is consistent with the innovation variance. We found the average estimated δ is about 1 when data are from a static system and less than 1 for dynamic synthetic data. Thus, at least in a simulation study the MDM was able to identify clearly the better system source on the basis of the lengths of data we record in experiments like these.

Markov equivalent DAGs

Note that, as we might expect, when data are static (the first column of Figure 4.4) it is difficult to distinguish the data generating DAGs. Using the model selection criteria proposed by West and Harrison (1997) where $-1 <$logBF< 1 suggests no significant difference amongst DAGs, there is no evidence for any particular DAG for almost 70% and 55% of the replications for $T = 100$ and $T = 200$, respectively. This percentage decreases to 38% for the largest sample size ($T = 300$). However, the equivalent DAGs remain indistinguishable (DAG3 and DAG4 were both selected for 56% of replications).

Another interesting result is that even when data follow a dynamic system but is fitted by a static model, the non-Markov equivalent DAGs are distinguishable whilst equivalent DAGs are not. For instance, when $\mathbf{W}^*(r) = 0.01\mathbf{I}_{p_r}$ and $T = 100$ (first row and third column), the value of LPL for DAG5 is smaller than the value for other DAGs, but there is no significant difference between the values of LPL for DAG3 and DAG4 when $\delta = 1$, which we could deduce anyway since these models are Markov equivalent (see *e.g.* Ali *et al.*, 2009). In contrast, there are important differences between the LPL of DAGs when dynamic data are fitted with dynamic models, DAG3 having the largest value of LPL.

Therefore, in the sense discussed above, the MDMs appear to select the appropriate direction of connectivity with a high success rate. However, their performance varies as a function of the innovation variance and sample size (note the distance between the lines of DAGs changes from one graph to another). For instance, as might be expected, the higher the sample size, the higher the chance of identifying the true DAG correctly. But T shows the largest impact in the results when the dynamics of the data are very slowly changing ($\mathbf{W}^*(r) = 0.001\mathbf{I}_{p_r}$). In this situation the percentage of replications in which the correct DAG was selected was 40%, 80% and 95%, for sample size equal to 100, 200 and 300, respectively.

An interesting result concerns different values of the innovation variance. When the connectivity does not change over time ($\mathbf{W}^*(r) = 0\mathbf{I}_{p_r}$), all three DAGs or the two Markov equivalent DAGs were selected for the overwhelming majority of replication (around 95%). However, there is a sharp increase in the performance of the model selection, from a model where $\mathbf{W}^*(r) = 0\mathbf{I}_{p_r}$ (static model) to a model where $\mathbf{W}^*(r) = 0.001\mathbf{I}_{p_r}$ and continues to improve as $\mathbf{W}^*(r) = 0.01\mathbf{I}_{p_r}$, with almost all of replications selecting the correct DAG. This demonstrates that the additional dynamic structures allow causal interactions to be identified

more clearly. However there is a slight decrease in the percentage of selected correct DAGs for $\mathbf{W}^*(r) = 0.1\mathbf{I}_{p_r}$ — around 95%. Perhaps this happens because it is not so easy to detect the correct network structure when the system has such a short memory.

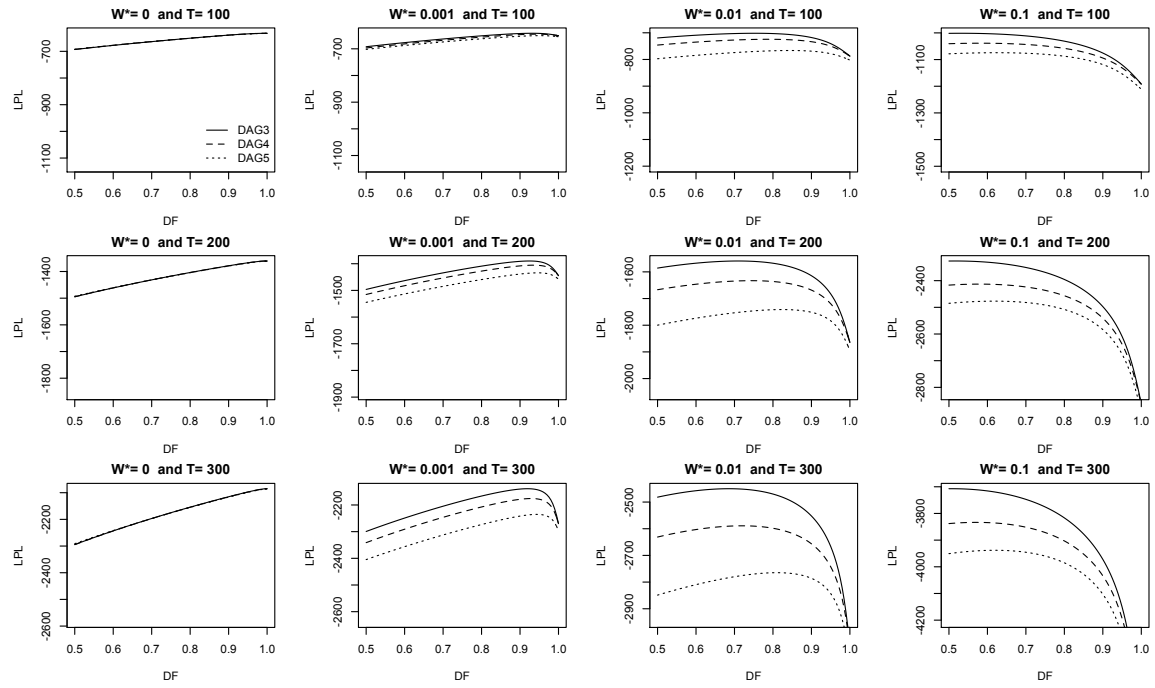


Figure 4.4: The average of log predictive likelihood over 100 replications versus different values of discount factor, for the true network DAG3 (solid lines), a Markov equivalent graph DAG4 (dashed lines) and a Markov non-equivalent graph DAG5 (dotted lines). The sample size increases from the first to the last row whilst the dynamic level (innovation variance) increases from the first to the last column. The range of the y-axis (LPL) has the same size of 500 for all graphs.

4.2 An Application of the MDM-IPA

4.2.1 A DCM Synthetic Study

In the last section, we provided the potential of the MDM to detect the true graphical structure using the simulated MDM data. We next analyse a synthetic dataset from a quite different model: the DCM fMRI forward model (see the description of this model in Section 2.3). In contrast to the previous simulations, this was supposed to simulate a real brain network. We chose the dataset *sim22* from Smith, S.M. *et al.* (2011), which has 5 regions, 10min-session, time resolution (*i.e.* sample rate) of 3.00s, 50 replications and the same graphical structure (see Figure 4.5). The connection strength was defined according to a random process and, therefore, varies over time, as Smith, S.M. *et al.* explain: "The

strength of connection between any two connected nodes is either unaffected, or reduced to zero, according to the state of a random external bistable process that is unique for that connection. The transition probabilities of this modulating input are set such that the mean duration of interrupted connections is around 30s, and the mean time of full connections is about 20s."

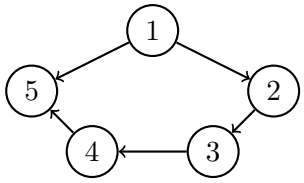


Figure 4.5: The graphical structure used by Smith, S. M. *et al.* (2011) to simulate data, using the DCM method.

Note that because the simulation was not driven by an MDM, we could not know a priori that this class of model would necessarily fit this dataset well. It, therefore, provided a much more rigorous test of our methods within the suite of the simulations available. We first compared the true DAG with Markov equivalent and Markov non-equivalent DAGs, and obtained the same result as considering the MDM synthetic data, in the previous section. We can see that, in general it was possible to detect the correct DAG, using a dynamic model (*i.e.* $\delta < 1$). Details of this analysis are given in Appendix B.1.

We will now discuss the performance of the MDM-IPA. The LPL was first computed for different values of discount factor δ, using a weakly informative prior with $n_0(r) = d_0(r) = 0.001$ and $\mathbf{C}_0^*(r) = 3\mathbf{I}_{p_r}$ for all r. The discount factors were chosen as the value that maximised the LPL, and so the average DF over all replications and nodes was around 0.85 (smaller than 1). Note that the MDM correctly identified that connectivities have been simulated to vary over time.

Smith, S. M. *et al.* (2011) compared different connectivity estimation methods ranging from the simplest approach which only considered pairwise relationships, such as correlation amongst the time series variables, to complex approaches which estimated a global network using all nodes simultaneously, such as BNs. The main measures that they used to compare these methods were *c-sensitivity* and *d-accuracy*. The former represents the ability of the method to correctly detect the presence of the connection, whilst the latter shows the ability of methods to distinguish the directionality of the relation between the nodes.

The first measure calculated was c-sensitivity as a function of the estimated strength

connectivity of the *true positive* (TP) edges which exist in both the true and the estimated graph, regardless the directionality, and the *false positive* (FP) edges that exist in the estimated graph but not in the true DAG. Here, we assess the performance of the methods in detecting the presence of a network connection, using the following measures:

- *Sensitivity* $= \#TP/(\#TP + \#FN)$, where # represents "the number of" and FN is an abbreviation for false negative edge which is a true connection that does not appear in the estimated graph. This measure represents the proportion of true connections which are correctly estimated;

- *Specificity* $= \#TN/(\#TN + \#FP)$, where TN is an abbreviation for a true negative edge which does not exist in both true and estimated graphs: *i.e.* the proportion of connections which are correctly estimated as nonexistent;

- *Positive Predictive Value* $= \#TP/(\#TP + \#FP)$: *i.e.* the proportion of estimated connections which are in fact true;

- *Negative Predictive Value* $= \#TN/(\#TN + \#FN)$: *i.e.* the proportion of connections estimated as nonexistent that do not exist in the true graph;

- *Success Rate* $= (\#TP + \#TN)/10$, where 10 is the total number of possible connections for an undirected graph with 5 nodes. This represents the proportion of correctly estimated connections.

The first row of Figure 4.6 shows the distribution of c-sensitivity calculated by Smith, S.M. *et al.* (2011; see the original paper for details) over 50 replications, for each modelling approach, and the blue line represents the mean of this distribution. In this plot, the smoothed histograms are reflected in the vertical axes and then have this violin form. According to this statistic, the best methods are algorithms that use Bayesian Network models. We therefore implemented two methods: the GES and the PC in the Tetrad IV[1]. The implementation of these methods is fairly easy, but unsurprisingly the computational time of the MDM is considerably higher than others because its descriptive search space is much larger. We estimated our sensitivity measures, as described above, and Figure 4.7 (left) shows the average of sensitivity measures over 50 replications for the MDM-IPA (blue bar), the GES (salmon

[1] `http://www.phil.cmu.edu/projects/tetrad/current.html`

bar) and the PC (green bar) methods. These approaches show satisfactory results for all measures, with a mean percentage above of 75%. Although the PC has the highest percentage in Specificity and Positive Predictive Value, the MDM performs better in the three other measures. For instance, the MDM correctly detected around 90% of the true connections whilst PC and GES detected about 75% (sensitivity measure). Moreover, the MDM has the highest overall percentage of correct connections (success rate).

As a second method of comparison, Smith, S.M. *et al.* (2011) evaluated the estimated connection strength of the true connectivity (i, j) subtracted by the estimated strength of the reverse connectivity (j, i), and the desired result would be a positive value. Figure 4.8 (left) shows the histogram for this *subtracted-Z* values over all true connections and over all replications for the MDM. It can be compared with the distributions shown in the second row of Figure 4.6. In addition, Smith, S.M. *et al.* (2011) proposed a way to compare the performance of the methods in detecting the *direction* of connectivity. The d-accuracy is calculated as the percentage of directed edges that are detected correctly. This measure is given in Figure 4.6, blue dots in the second row, and in Figure 4.7 (right). Again the MDM obtained some of the best results for this measure. Other methods that also had good results according to this criterion were Patel's measures (Patel *et al.*, 2006) and Generalised synchronization (Gen Synch; Quian Quiroga *et al.*, 2002). We note that the performance of LiNGAM was poor when compared with other methods.

Although the d-accuracy of Patel's τ and Gen Synch is not substantially different from that of the MDM (Figure 4.7, right), these two former methods have only moderate c-sensitivity scores (Smith, S.M. *et al.*, 2011). The opposite pattern can be seen for the methods based on the BN: they perform well in c-sensitivity but poorly in d-accuracy. Thus, only the MDM performed well in all the measures at the individual level of analysis.

We suggest another criterion for assessing the directionality of the edge: using the logBF criterion. Figure 4.8 (right) shows the histogram of the logBF comparing the model with true connection to the model with the reverse connection. For most of the edges and replications, the logBF is positive, showing more evidence in favour of the DAG with right graphical structure. Details about this analysis are described in Appendix B.2.

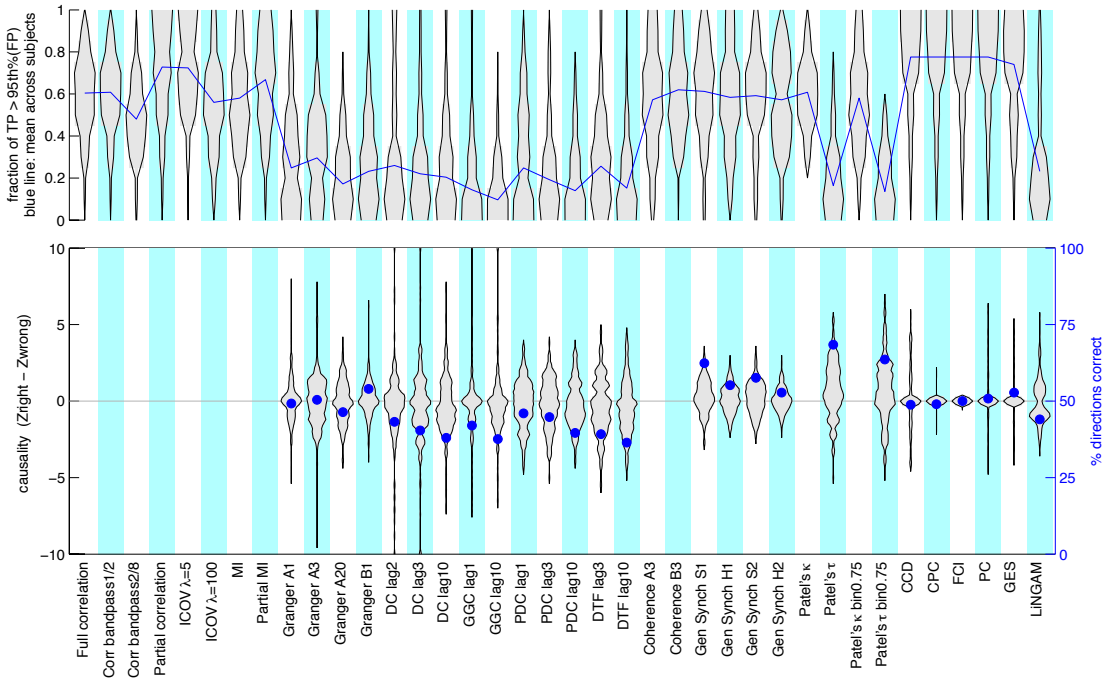


Figure 4.6: Results for simulation 22. The above and below figures provides the c-sensitivity and the d-accuracy measures, respectively (unpublished result from Smith, S. M. *et al.*, 2011).

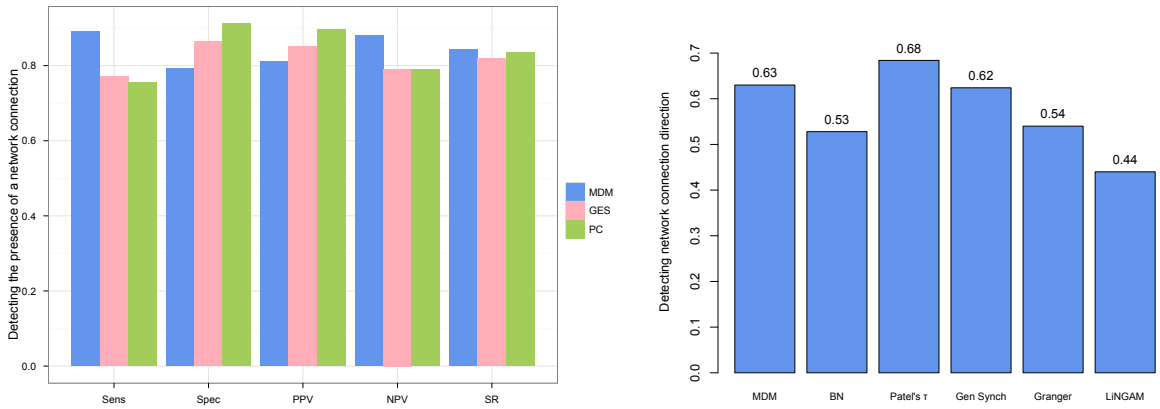


Figure 4.7: (*left*) The average over 50 replications of the *sensitivity* (Sens) $= TP/(TP+FN)$; *specificity* (Spec) $= TN/(TN+FP)$; *positive predictive value* (PPV) $= TP/(TP+FP)$; *negative predictive value* (NPV) $= TN/(TN+FN)$; (SR) *success rate* $= (TP+TN)/$(total number of connections) for three methods: MDM (blue bar), GES (salmon bar) and PC (green bar). (*right*) The average over 50 replications of the percentage of directed connections that was detected correctly for some methods. The results of this second figure are from Smith, S.M. *et al.* (2011), except for the method MDM.

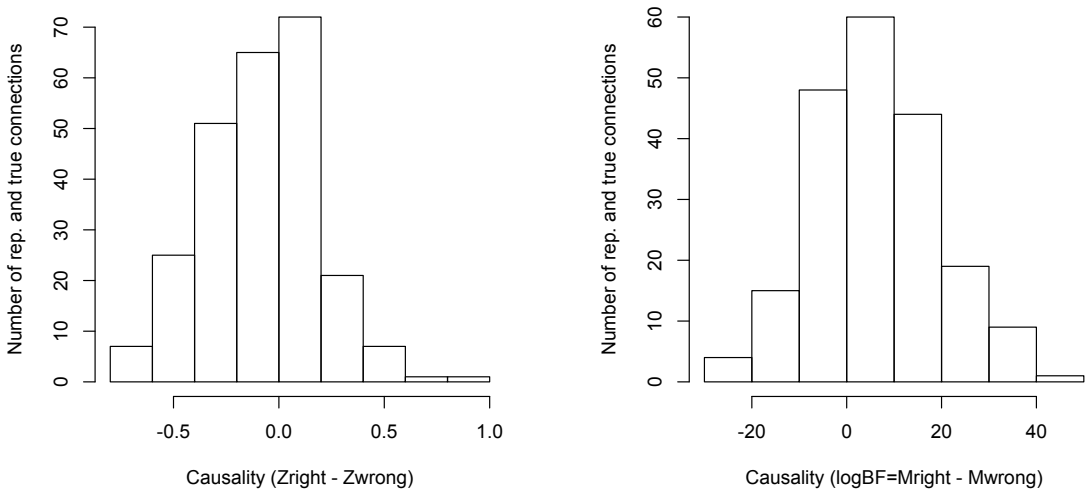


Figure 4.8: (*left*) The histogram of subtracted-Z defined as the estimated connection strength of the true connectivity subtracted by the estimated strength of the reverse connectivity over all true connections and over all replications. (*right*) The histogram of logBF comparing the model with true connection to the model with the reverse connection, considering only nodes connected by particular edges over all possible comparisons and all replications.

4.2.2 An MDM Synthetic Study

We have shown the performance of the MDM-IPA considering synthetic data from 5-node networks. It would be interesting to see how this search algorithm performs with a larger number of nodes. However, although Smith, S.M. *et al.* (2011) provided higher dimensional DCM fMRI synthetic data, none of these was generated considering the stochastic process in connectivity strengths. We, therefore, generated MDM data based on our analysis of the resting-state experiment studied in Section 6.3. More specifically, we fixed the 11 nodes and 230 time points in this experiment and simulated data of this size assuming as true to the best fitting MDM we found for the original data set (Figure 4.9 (a)). More details about the simulation process are described as follows.

The initial values for the regression parameters were defined as the average of estimated values over time from the real data, *i.e.* zero for intercept parameters, 0.25 for the connection $2 \rightarrow 4$, 0.18 for the connection $8 \rightarrow 4$, 0.50 for the connection $3 \rightarrow 5$, 0.80 for the connection $7 \rightarrow 6$, 0.39 for the connection $8 \rightarrow 7$, and 0.65 for the connection $10 \rightarrow 9$. The observational variance was also defined considering the estimated variance of variables from the real data, *i.e.* 0.010, 0.191, 0.036, 0.005, 0.018, 0.011, 0.010, 0.006, 0.016, 0.014 and

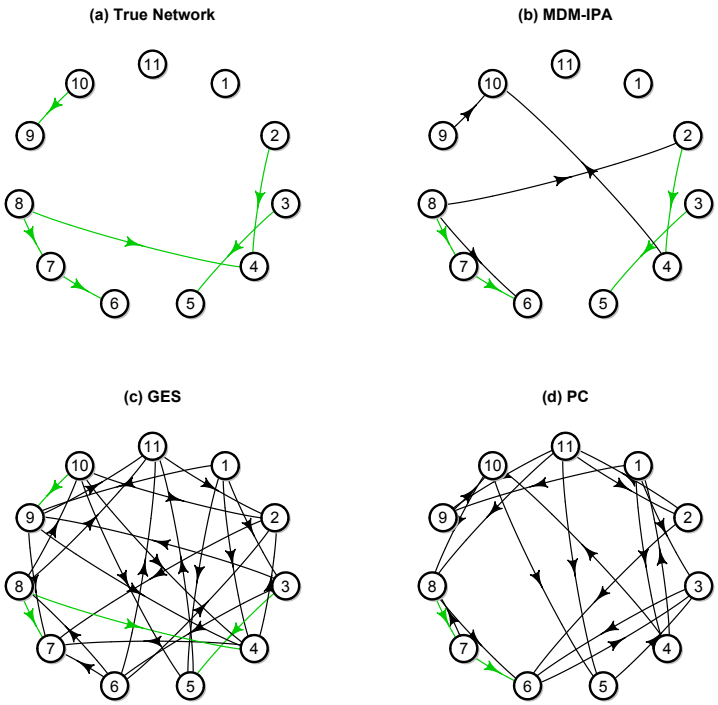


Figure 4.9: The MDM synthetic data was generated considering (a) true network. This is an example of estimated network considering (b) MDM-IPA, (c) GES and (d) PC algorithms, for a particular replication. True edges are in green.

0.013 for the variables of nodes 1 to 11, respectively. Thus we set

$$\theta_{ti}^{(k)}(r) = \theta_{t-1i}^{(k)}(r) + w_{ti}^{(k)}(r), \quad w_{ti}^{(k)}(r) \sim \mathcal{N}(0, W^{(k)}(r)),$$

for $r = 1, \ldots, 11$; $t = 1, \ldots, 230$; $i = 1, \ldots, 50$ replications (the same as the last section); $k = 1, \ldots, p_r$; $p_r = 1$, for $r \in \{1, 2, 3, 8, 10, 11\}$; $p_r = 2$, for $r \in \{5, 6, 7, 9\}$; $p_4 = 3$; $W^{(k)}(r) = W^{*(k)}(r) \times V(r)$ and $W^{*(k)}(r)$ is the k^{th} element of the diagonal of matrix $\mathbf{W}^*(r) = 0.05\mathbf{I}_{p_r}$. Observed values were then simulated using the following equations:

$$\begin{aligned}
Y_{ti}(j) &= \theta_{ti}^{(1)}(j) + v_{ti}(j); \\
Y_{ti}(4) &= \theta_{ti}^{(1)}(4) + \theta_{ti}^{(2)}(4)Y_{ti}(2) + \theta_{ti}^{(3)}(4)Y_{ti}(8) + v_{ti}(4); \\
Y_{ti}(5) &= \theta_{ti}^{(1)}(5) + \theta_{ti}^{(2)}(5)Y_{ti}(3) + v_{ti}(5); \\
Y_{ti}(7) &= \theta_{ti}^{(1)}(7) + \theta_{ti}^{(2)}(7)Y_{ti}(8) + v_{ti}(7); \\
Y_{ti}(6) &= \theta_{ti}^{(1)}(6) + \theta_{ti}^{(2)}(6)Y_{ti}(7) + v_{ti}(6); \\
Y_{ti}(9) &= \theta_{ti}^{(1)}(9) + \theta_{ti}^{(2)}(9)Y_{ti}(10) + v_{ti}(9);
\end{aligned}$$

where $j \in \{1, 2, 3, 8, 10, 11\}$, $v_{ti}(r) \sim \mathcal{N}(0, V(r))$, and other parameters were defined as before.

Algorithms GES, PC and MDM-IPA were then applied for 50 replications. Using a weakly informative prior for the MDM and considering the network estimated by the MDM-IPA, the average DF over replications was 0.83. This is very close to the one found in 5-node networks study (0.85). In this sense, both sets of synthetic data (the DCM with 5 nodes and the MDM with 11 nodes) have a similar variability of connections over time. When Smith, S.M. *et al.* (2011) compared the results of data over different numbers of nodes, they concluded that the classification order of methods considering c-sensitivity and d-accuracy measures is extremely similar for 5, 10, and 15 nodes. Here, considering the algorithms GES, PC and MDM-IPA, we came to the same conclusion. Table 4.1 shows that in general the MDM-IPA has the highest c-sensitivity measures and much better scores. While almost 80% of the estimated connections are actually true (PPV measure) for the MDM-IPA in both sets of synthetic data, for the GES this percentage was 85% in 5-node networks and it decreased to around 20% in 11-node networks (the PC provided a similar pattern). A plausible reason for this is that the number of false positive connections is dramatically higher for the GES and the PC than for the MDM-IPA in 11-node networks, *i.e.* the average $\#FP$ over replications for the GES, the PC and the MDM-IPA, respectively, was around 0.7, 0.4 and 1.0 in 5-node networks, and around 10, 18 and 2 in 11-node networks (see examples of estimated graphs in Figure 4.9 (b), (c) and (d)). Although the prevalence of FP increases with the number of nodes, their connectivity strengths usually are close to zero, as shown below.

c-sensitivity	5-node-networks			11-node-networks		
measures	MDM-IPA	GES	PC	MDM-IPA	GES	PC
Sens	0.89	0.77	0.76	0.83	0.85	0.74
Spec	0.79	0.86	0.91	0.97	0.64	0.80
PPV	0.81	0.85	0.90	0.82	0.23	0.32
NPV	0.88	0.79	0.79	0.98	0.97	0.96
SR	0.84	0.82	0.83	0.96	0.67	0.79

Table 4.1: The average over 50 replications of the *sensitivity* (Sens) $= TP/(TP + FN)$; *specificity* (Spec) $= TN/(TN + FP)$; *positive predictive value* (PPV) $= TP/(TP + FP)$; *negative predictive value* (NPV) $= TN/(TN + FN)$; (SR) *success rate* $= (TP + TN)/$(total number of connections) for three methods: the MDM-IPA, the GES and the PC, considering 5-node-network (also shown in Figure 4.7, left) and 11-node-network synthetic data.

When we focused on the d-accuracy criteria, the MDM-IPA also demonstrated greater

power in detecting the direction of connectivity than the GES and the PC — 60%, 45% and 26% of directed edges were detected correctly in 11-node networks for the MDM-IPA, the GES and the PC, respectively.

In addition, we evaluated the proportion of time that the true value of connection i for node r is inside the 95% smoothed HPD intervals. Thus, we calculated

$$PT_{ri} = \frac{\sum_{t=1}^{T} I(\theta_t^{(i)}(r))}{T},$$

where $I(\theta_t^{(i)}(r)) = 1$, if the true value of $\theta_t^{(i)}(r)$ is inside the 95% smoothed HPD interval, and it is zero otherwise. The average of PT_{ri} over all replications, considering only TP connections (green edges in Figure 4.9 (a)), turned out to be 96%. The MDM-IPA thus appeared to be efficient not only in detecting the edges, but also in estimating the connectivity strengths. (see *e.g.* Figure 4.10, connection $3 \rightarrow 5$).

There are two kinds of FP connections: (*situation 1*) the edges exist in both the true and the estimated networks, but with opposite directions (see *e.g.* the connection $9 \rightarrow 10$ in Figure 4.9 (b)), and (*situation 2*) the edges exist only in the estimated network (see *e.g.* the connection $4 \rightarrow 10$ in Figure 4.9 (b)). Considering the true value of these FP regression parameters as zero, the average PT_{ri} over all replications, considering the FP connections with opposite directions in the true network (*situation 1*) was about 30% (see *e.g.* Figure 4.10, the connection $9 \rightarrow 10$). This is no surprise, when the opposite connection strength is higher than zero. In contrast, the average PT_{ri} over all replications, considering the FP connections that do not exist even on an undirected true network (*situation 2*) was 75%. This shows that when the MDM-IPA provides spurious connections, the associated connectivity strengths are usually non-significant (see *e.g.* Figure 4.10, the connection $8 \rightarrow 2$).

It is interesting to note that this appearance of spurious but weak dependences is a well known phenomenon when fitting more standard graphical models using Bayes factor methods. More robust conjugate Bayes model selection methods have recently been investigated using non-local priors (see *e.g.* Consonni and La Rocca, 2010), and analyses of these could provide promising alternatives to the scoring methods in this thesis (see Section 7.1).

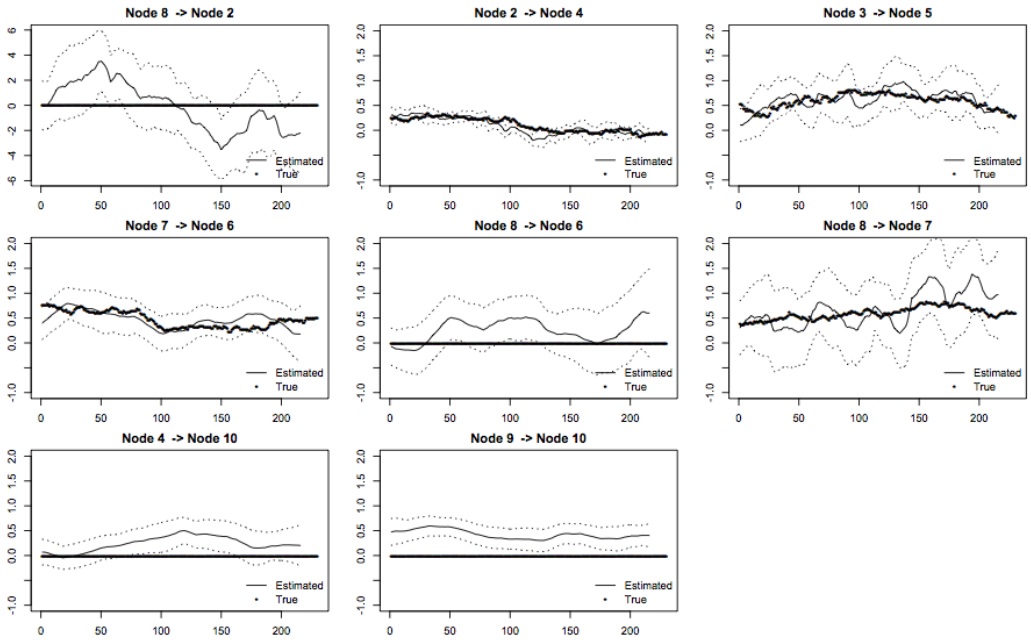


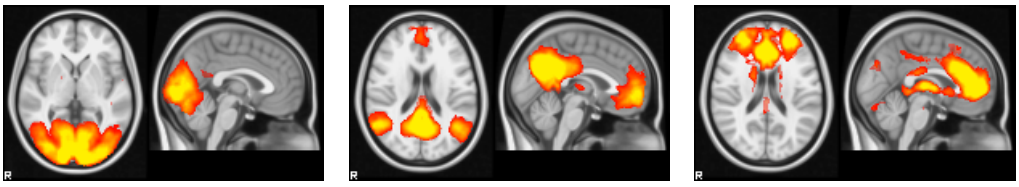
Figure 4.10: The smoothed posterior mean (solid lines) of connections found for a particular replication using the MDM-IPA (Figure 4.9 (b)) and their 95% HPD intervals (dotted lines). The stars are the true values of connectivity strengths.

4.3 The ICA Data Analysis

This study focuses on real fMRI time series data, where 36 healthy adults at rest were observed over six minute intervals (Smith *et al.*, 2009). The first step in the study of fMRI connectivity is data reduction. This is as usually achieved by summarizing the data as a set of time series derived from predefined regions of interest (ROIs). The use of ROIs has various drawbacks: the choice of ROI set to use is somewhat arbitrary, poorly chosen ROIs may mix heterogeneous brain regions, and, further, an ROI based network cannot easily represent spatially overlapping networks (Smith *et al.*, 2011). An alternative is to use Independent Components Analysis (ICA). ICA generates data-driven spatial patterns that describe the structure of local and long-range connectivity, which addresses a number of the problems with ROIs (see Section 2.3).

Here we have used such an ICA approach. Each subject's image data was transformed into a standard atlas space, and then all subjects' data concatenated temporally. After an initial principal component data reduction to 20 components, ICA produced a set of 20 matched pairs of spatial components (that are orthogonal and maximally statistical independent) and temporal loadings (that may be correlated). Of these 20, 10 spatial components were identi-

fied as well-known resting-state networks (Smith *et al.*, 2009). The temporal loadings, split back into 36 separate time series, of length 176 each, express the temporal evolution of the corresponding spatial pattern in each subject and were the source of the data for our MDM modelling. To demonstrate the methodology developed in Chapter 3, we considered here the networks of just three of these components. Henceforth we refer to those components as *regions* (to distinguish them from the network we build between these three regions): Region 1, a visual network composed of medial, occipital pole and lateral visual areas (comprised of the average of the 3 visual networks in Smith *et al.*, 2009); Region 2, the "default mode network" (DMN) comprising posterior cingulate, bilateral inferior-lateral-parietal and ventromedial frontal areas; and Region 3, an "executive control" network that covers several medial-frontal areas including anterior cingulate and paracingulate cortex (Figure 4.11). We randomly selected subject 19 to analyse.



(a) Visual (b) DMN (c) Executive Control

Figure 4.11: Three networks from independent component analysis of the 36-subject resting fMRI dataset (Smith *et al.*, 2009): Region 1 - Visual (a), Region 2 - Default Mode Network (b) and Region 3 - Executive Control (c).

A simple preliminary analysis of this data clearly demonstrates the scientifically predicted dynamically evolving changes in strength of dependency between these series. For example plots of estimates of regression coefficients are given in Figure 4.12, using simple regression models, fitted to estimate a linear relationship between each pair of regions and based on a moving window of 30 time points. The plots clearly exhibit some large drifts in dependency strengths over time. This strongly suggests that a time-varying flexible model such as a simple linear MDM should be used for this application.

The MDMs were fitted using a weakly informative prior and the discount factor δ was estimated for each region and each graphical structure. Table 4.2 shows the LPL scores calculated for all possible sets of parents per region. The best scoring model has the variables associated with the Regions 1 and 2 with no parents and the Region 3 with the other two

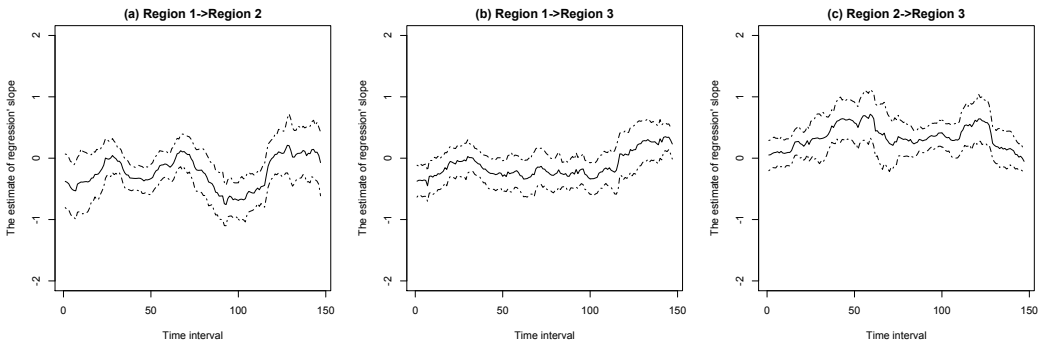


Figure 4.12: The posterior mean and 95% HPD intervals for regression parameters between every 2 regions: (a) *Visual* → *DMN*; (b) *Visual* → *Executive Control*; (c) *DMN* → *Executive Control* over 147 time intervals.

Region	**Parent**	**Score**
1 - Visual	No	-418.09
	2	-443.66
	3	-451.20
	2 and 3	-455.52
2 - DMN	No	-421.93
	1	-444.68
	3	-443.57
	1 and 3	-444.39
3 - Executive Control	No	-414.11
	1	-412.40
	2	-407.16
	1 and 2	-407.09

Table 4.2: Evidence for each region under all possible sets of parents. Score was calculated as $\text{LPL}[Y(r)|Pa(r)]$. The score for a particular network is calculated as $\text{LPL} = \text{LPL}[Y(1)|Pa(1)] + \text{LPL}[Y(2)|Pa(2)] + \text{LPL}[Y(3)|Pa(3)]$. The higher score the higher evidence for this particular model.

regions as its parents, see Figure 4.13. We found small values of the discount factor for Regions 1 and 2, *i.e.* they have the shortest memory. This is scientifically plausible, as a region not driven by external stimuli may indeed be expected to have the noisiest signal. This result was also found for other datasets, as shown below. For Region 3, executive control, the DF was found as almost 0.95.

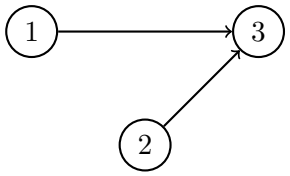


Figure 4.13: ICA-DAG: The best DAG that was selected according to scores in Table 4.2 — visual region (node 1) and DMN (node 2) are the parents of executive control (node 3).

We next use this dataset to illustrate the use of parent-child Monitor. Suppose we want to confirm the relations "parent-child" for Region 3. Figure 4.14 provides estimates of connectivities over time. The connectivity from the visual region to the executive control (Figure 4.14(a)) seems not to be significant in the second half of the series. The significance of this connectivity is reflected in:

$$\text{logBF}_{31} \quad = \quad \log p(\mathbf{y}(3)|\mathbf{y}(1), \mathbf{y}(2)) - \log p(\mathbf{y}(3)|\mathbf{y}(2)).$$

Figure 4.15 provides individual and cumulative logBF for each time (the first 15 points were disregarded as burn-in), comparing visual region and DMN as the parents of executive control (ICA-DAG) with only DMN influences the executive control region. The cumulative logBF shows strong evidence for ICA-DAG, which has both visual region and DMN as the parents of executive control region, at the beginning of the series.

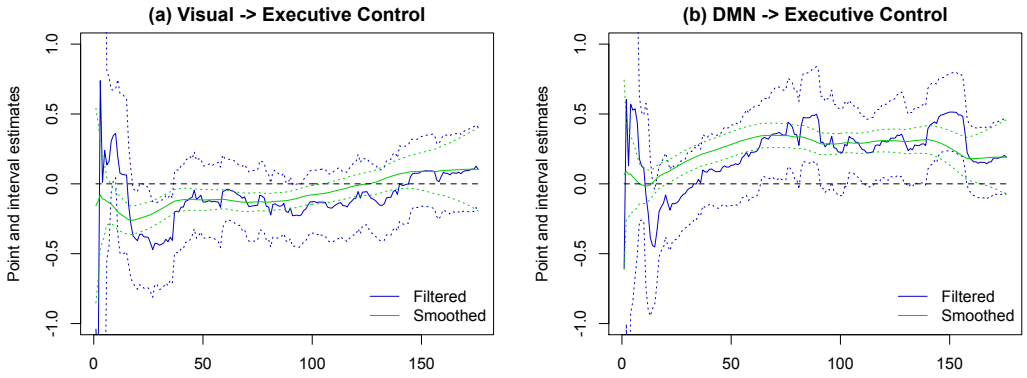


Figure 4.14: The filtered (blue) and smoothed (green) posterior mean (solid lines) and 95% HPD interval (dotted lines) for regression parameters (a) $\theta^{(2)}(3)$ - connectivity $1 \rightarrow 3$, visual region influences executive control region, and (b) $\theta^{(2)}(3)$ - connectivity $2 \rightarrow 3$, DMN influences executive control region.

Observe now the connectivity from DMN to the executive control in the Figure 4.14(b). There is a sharp increase in the strength of this connectivity in the first quarter of the series, and the connectivity remains significant subsequently. This considerable variation in the connection strengths is consistent with other scientific reports on the non-stationarity of resting-state fMRI (Ge *et al.*, 2009; Allen *et al.*, 2012; Leonardi *et al.*, 2013) and further demonstrates the plausibility of this class of model to capture real scientific phenomena. One possible explanation for the observed apparent changes in connectivity strengths proposed by Chang and Glover (2010) is that the level of attention, arousal and

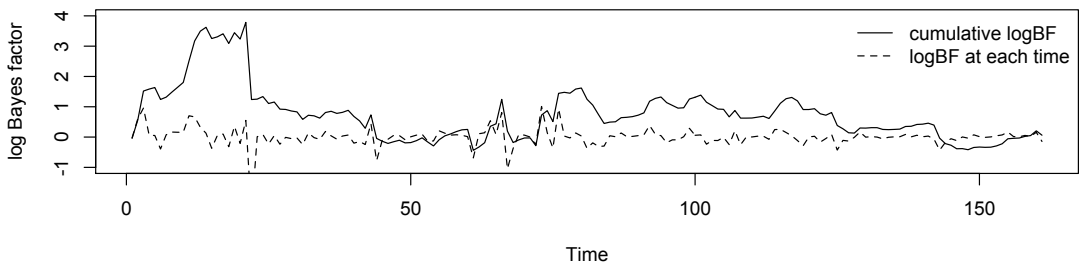


Figure 4.15: The logBF at each time (dashed lines) and cumulative logBF (solid lines) comparing visual region and DMN as the parents of executive control (IPA-DAG) with only DMN influences the executive control region.

daydreaming can differ during the resting-state experiment, and this is reflected through the measurements.

The influence of individual observations on the model analysis can be verified using the method shown in Section 3.5.3. Recall that $K_t(r)$ is the Kullback-Leibler distance between the joint posterior distribution and the jackknifed posterior distribution of parameters $\boldsymbol{\theta}(r)$ and $\phi(r)$. This measure is evaluated as $I_t(r) + J_t(r)$, where $I_t(r)$ and $J_t(r)$ represent the differences between the posterior and the jackknifed distributions of connections and observational variances, respectively. Figure 4.16 shows this influence measure $K_t(r)$ plotted against t. The two observations of executive control region, times 5 and 158, seem to have a higher effect on the estimation of the parameters than other time points (Figure 4.16 (c)). We noted similar pattern of $K_t(r)$ for the other measures $I_t(r)$ and $J_t(r)$. In the next section, we will show more illustrations of the use of node monitor and global monitor to help make scientific deductions about real fMRI datasets.

4.4 A 4-node Resting-State FMRI Data Analysis

Our second real data consist of 197 fMRI resting-state time-points (TR=2s) for 4 regions of interest: Region 1 - Posterior Cingulate (PC); Region 2 - Anterior Frontal (AF); Region 3 - Left Lateral Parietal (LP) and Region 4 - Right Lateral Parietal (RP). There is information for three sessions for each one of 25 subjects, and so there are 75 datasets. Firstly four different graphical structures were chosen for representing the scientific beliefs about the brain connectivities (Figure 4.17). RS-DAG1 represents the idea that Posterior Cingulate

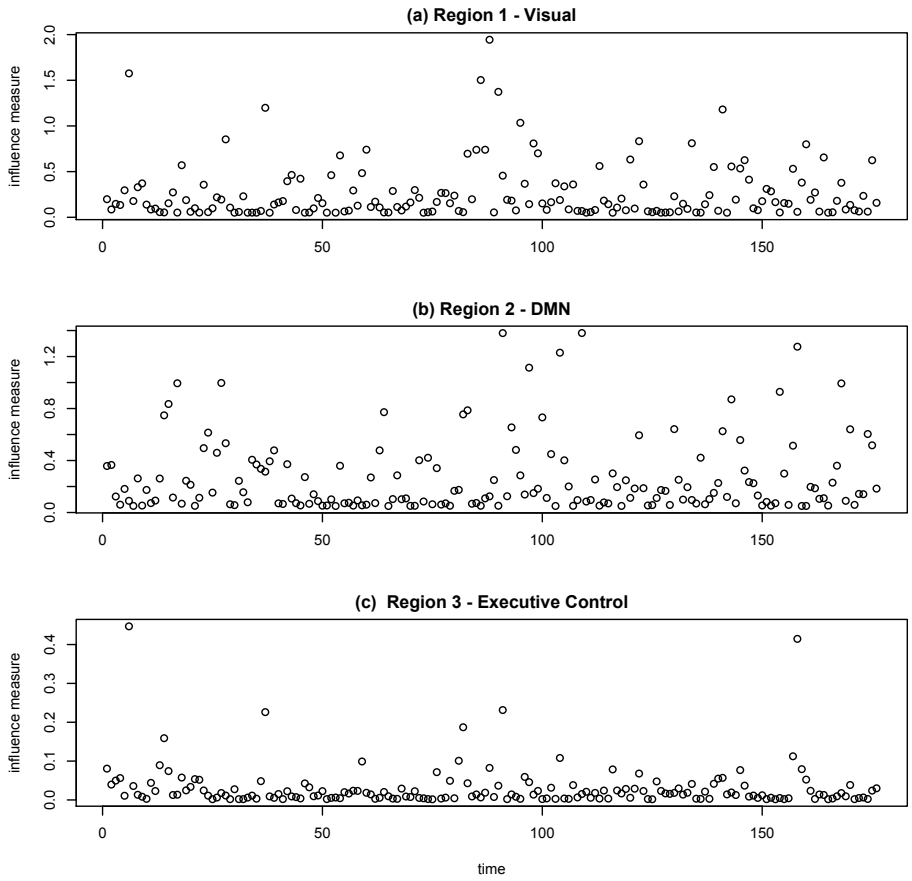


Figure 4.16: Overall influence measures $K_t(r)$ over time for every region.

hub drives other regions whilst RS-DAG2 means that Posterior Cingulate hub driven by Anterior Frontal and Left and Right Lateral Parietal. In RS-DAG3, the information flows in a forward way while, in RS-DAG4, the information flows in a backward way. Forthwith the discount factor was chosen for each region, considering one DAG and one particular dataset (see the average of δ across datasets in Figure 4.18). As discussed in the previous section, the regions not having parents show a more dynamic behaviour (smaller δ), *e.g. AF* in RS-DAG2. The DAG that maximises the log predictive likelihood was selected for each session and each subject. The RS-DAG4 was chosen for most of datasets (54.7%), following by RS-DAG1 (41.3%). We noted that, in general, 91% of the sessions of the same subject have the same result.

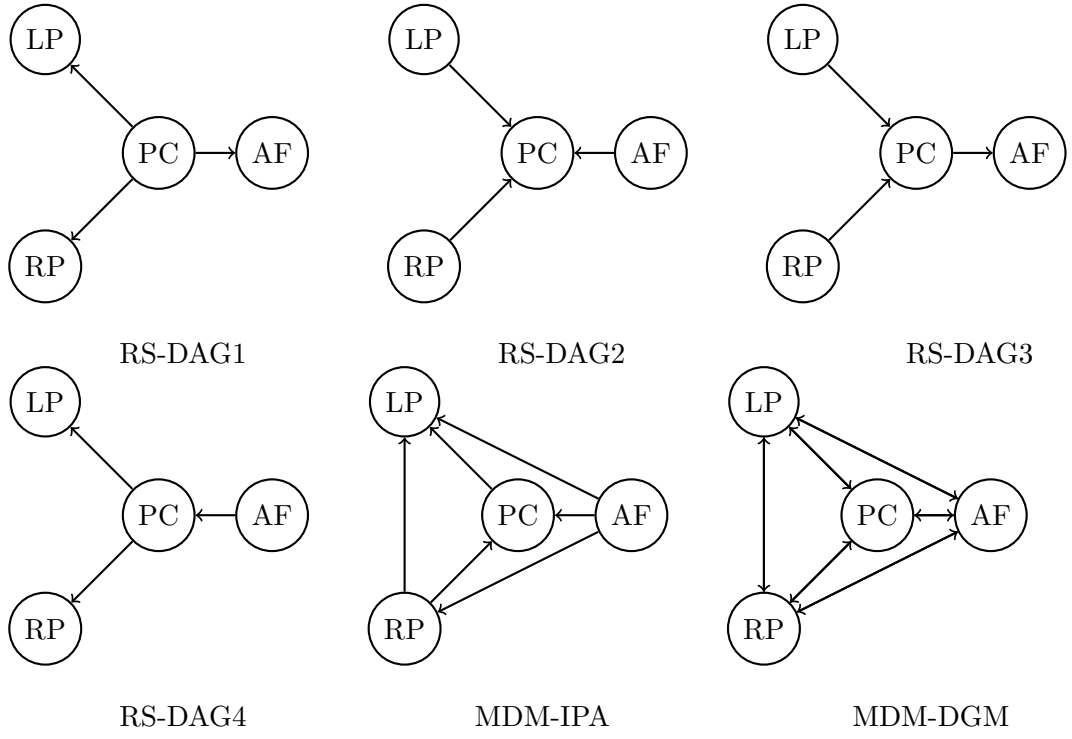


Figure 4.17: The graphical structures from RS-DAG1 to RS-DAG4 were used in the first learning process. The RS-DAG4 was chosen for most of datasets (54.7%), following by RS-DAG1 for 41.3% of runs. Then the scores were summed over all datasets and the MDM-IPA and the MDM-DGM were applied. PC means the posterior cingulate area (node 1), AF means the anterior frontal area (node 2), LP means the left lateral parietal area (node 3) and RP means the right lateral parietal area (node 4).

We now can give further illustrations of the use of diagnostic monitors developed in Section 3.5. Here we selected the subject 22 and session 2 because its chosen graphical structure was RS-DAG4, and so in this sense it was a typical experimental subject. The global monitor is applied considering three DAGs: RS-DAG1 and RS-DAG4 are *Markov*

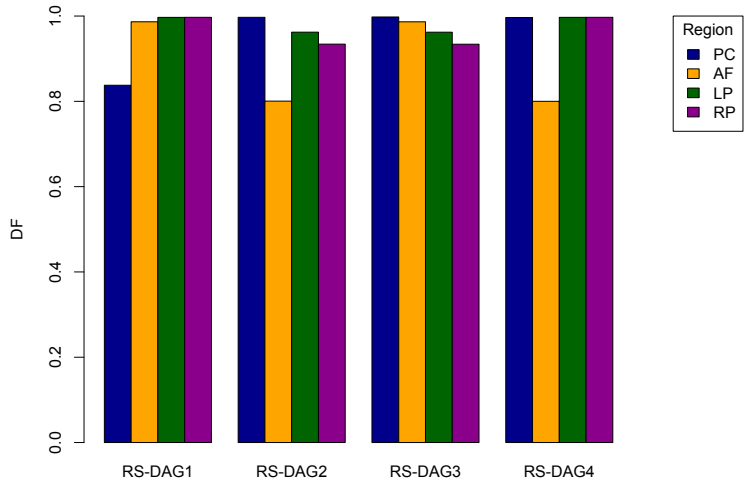


Figure 4.18: The average of parameter δ across 75 datasets (3 sessions for each 25 subjects) for each DAG and region.

equivalent graphs while both are *Markov non-equivalent* to RS-DAG3. Figure 4.19 shows the LPL for different values of the discount factor (note that in this figure we use the same value of δ for all nodes). Recall that the measure of model selection, logBF, is calculated as the difference between LPL of two DAGs, and so the higher distance between the two lines of the DAGs in the Figure 4.19, the higher evidence for the DAG with larger LPL. Therefore, RS-DAG4 should be chosen for all values of the discount factor, except for $\delta = 1$ when the LPL is approximately the same for RS-DAG1 and RS-DAG4. Thus, considering BN, it is not possible to distinguish between these two equivalent DAGs although it is clear the difference between non-equivalent DAGs (note that the RS-DAG3 has the smallest LPL even when $\delta = 1$). Although we have shown this result using synthetic data in Section 4.1, here we emphasise this characteristic of MDM in detecting causality compare two Markov equivalent DAGs using real dataset.

It is possible to distinguish Markov equivalent graphs using the MDM because the system equation allows that causal relations are estimated by the past connection information (Queen and Albers, 2009). Thus, although the relation between variables is contemporaneous, the model selection is updated over time, based on the posterior probability of model M using the recurrence

$$p(M|\mathbf{y}^t) \propto p(M|\mathbf{y}^{t-1})p(\mathbf{y}_t|M, \mathbf{y}^{t-1}).$$

Thus two Markov equivalent DAGs, say M_1 and M_2, cannot only be distinguished using the rate $p(\mathbf{y}_t|M_1, \mathbf{y}^{t-1})/p(\mathbf{y}_t|M_2, \mathbf{y}^{t-1})$ at a particular time t, but also using the rate $p(M_1|\mathbf{y}^T)/p(M_2|\mathbf{y}^T)$, as show before using BF. For instance, Figure 4.20 shows the cumulative log Bayes factor comparing two Markov non-equivalent graphs, RS-DAG4 with RS-DAG3 (orange lines), using a static model (dotted lines) and dynamic models (solid lines). In any case, it was possible to distinguish these two graphs. In contrast, the model selection measure did not show evidence for a particular graph when RS-DAG4 was compared to a Markov equivalent graph (RS-DAG1), in a static model (blue dotted line). But, using a dynamic model (blue solid line), the evidence to RS-DAG4 increases over time.

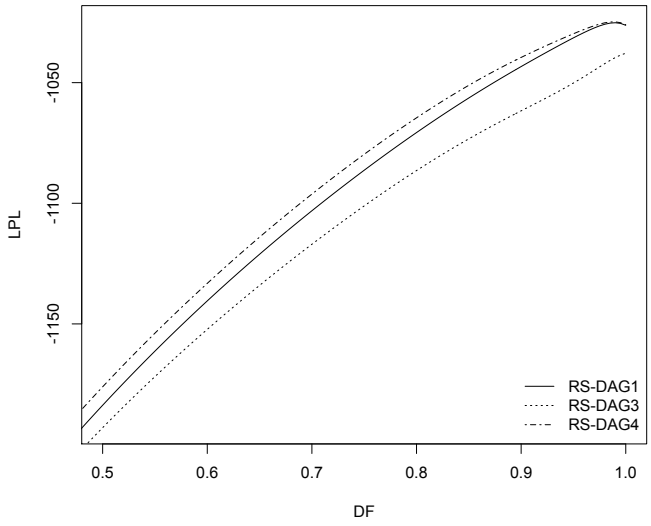


Figure 4.19: The log predictive likelihood versus different values of the discount factor (DF). RS-DAG1 (solid line) and RS-DAG4 (dotdashed line) are considered Markov equivalent whilst neither are equivalent to RS-DAG3 (dotted line).

As discussed above, a simple LMDM can easily be embellished in order to solve problems detected by diagnostic measures. For example, Figure 4.21 shows the time series, ACF and the cumulative sum plot of the standardized conditional one-step forecast errors for each region. Note that the ACF-plot suggests autocorrelation at lag 4 and 2 for Regions PC and RP, respectively (PACF-plot shows similar pattern of ACF-plot). This feature can still be modelled within the MDM class by making a local modification. For example, the past of the Regions PC and RP may be included in their observation equations. That is,

$$\begin{aligned} Y_t(1) &= \theta_t^{(1)}(1) + \theta_t^{(2)}(1)Y_t(2) + \theta_t^{(3)}(1)Y_{t-4}(1) + v_t(1); \\ Y_t(4) &= \theta_t^{(1)}(4) + \theta_t^{(2)}(4)Y_t(1) + \theta_t^{(3)}(4)Y_{t-2}(4) + v_t(4). \end{aligned}$$

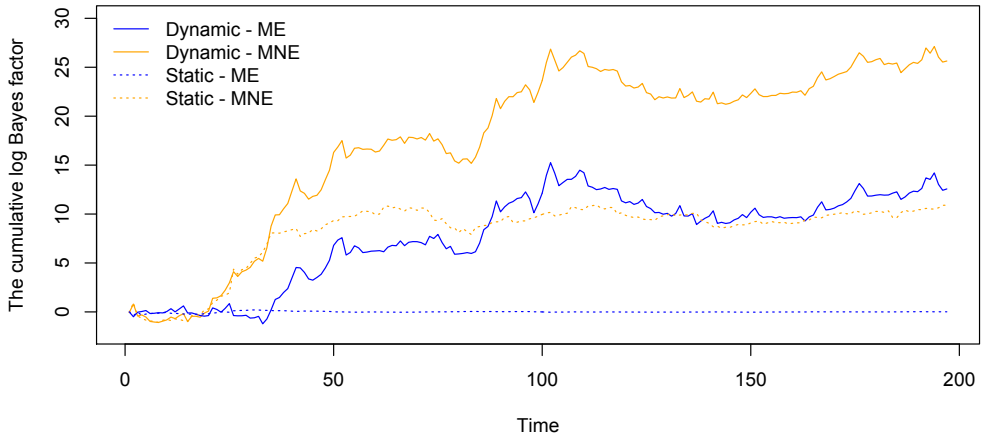


Figure 4.20: The cumulative log Bayes factor comparing RS-DAG4 to a Markov equivalent graph RS-DAG1 (blue lines), and comparing RS-DAG4 to a Markov non-equivalent graph RS-DAG3 (orange lines), considering a static model ($\delta = 1$; dotted lines) and a dynamic model ($\delta < 1$; solid lines).

Figure 4.22 provides the residual analysis plots considering the model with a lag for Regions PC (first column) and RP (third column). We can see that the insertion of the past of the observation variable improves the ACF-plot. However, the cumulative sum of forecast errors (first column and third row) exhibits a non-random pattern, which suggests an additional feature: the presence of change points for Region 1 - PC. In Section 3.5.3 we described a simple method to model this phenomenon as follows. Firstly, the logBF or the cumulative logBF is calculated in each time point comparing two models. If this measure is less than a particular threshold, a new model is fitted which entertains the possibility that a change point may have occurred.

Adopting this monitoring algorithm provided in Section 3.5.3 and comparing the current graph with the graph where there is no parent from Region PC, with a threshold of -1.6, four time points were suggested as change points. It was straightforward to run a new MDM with a change point at the identified point, simply by increasing the state variance of the corresponding system error at these two points. Figure 4.23 shows these change points (vertical dashed lines) and the filtered posterior estimates for all connectivities, considering three models: original RS-DAG4 (blue lines), the MDM with lags 4 and 2 for Regions 1 and 4, respectively (green lines), and the MDM with lags, as before, and change points for Region 1 (orange lines).

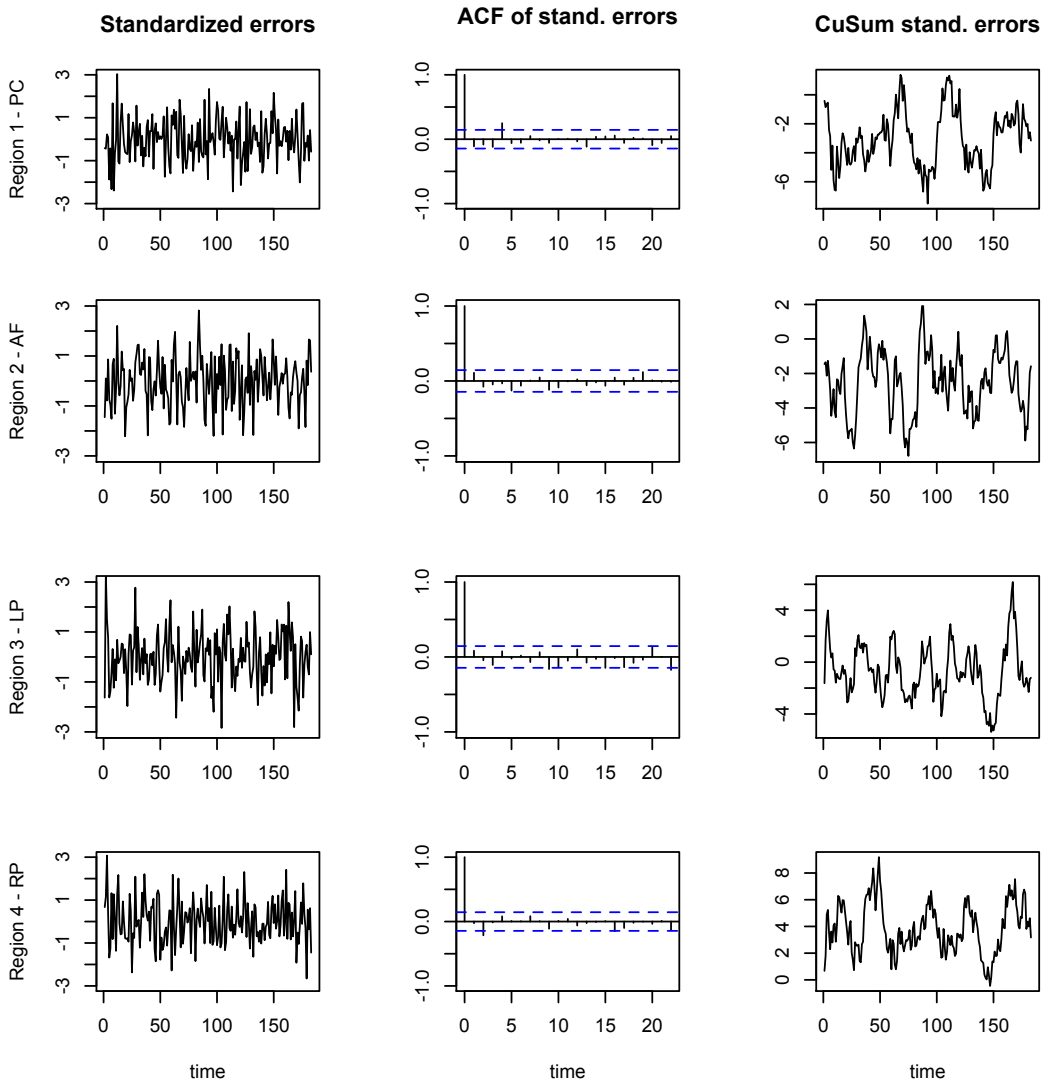


Figure 4.21: The time series (first column), ACF (second column) and the cumulative sum plot (third column) of the standardized conditional one-step forecast errors for each region (one region per row). This illustrates the use of the node monitor.

Note that there is a dramatic decrease in the precision of filtered posterior estimate of connectivity AF → PC, when the change point is detected (Figure 4.23 (a)). As discussed before, this large innovation variance allows the model adapts to the new data. However the precision increases after few time points, because the impact of this large prior variance declines over time. In addition, Figure 4.23 shows that the embellishment in the models of Regions 1 and 4 only affects the estimate of connectivity that indicates these regions as child. For instance, the estimate of connectivity PC → RP changed (from blue to green line) when the past of variable $Y_t(4)$ was included in the model for Region 4 - RP, but it did not change when the monitoring process was used for Region 1 - PC (Figure 4.23 (c), orange is the same as green line).

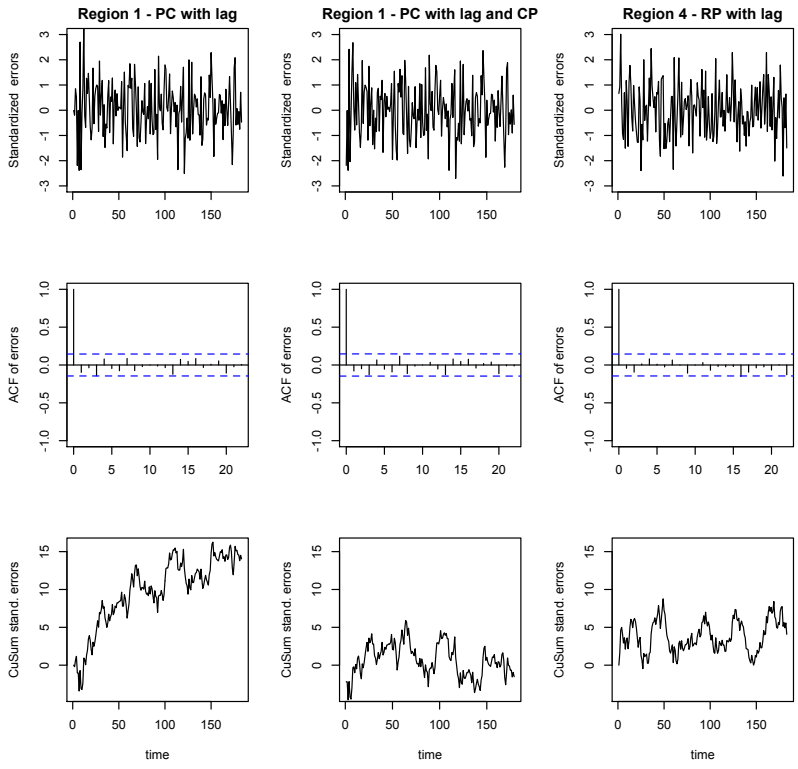


Figure 4.22: The time series, ACF and cumulative sum plot of the standardized conditional one-step forecast errors for Region1 - PC considering lag 4 (first column), for Region1 - PC considering lag 4 and change points (second column) and Region 4 - RP considering lag 2 (third column).

Normality, heteroscedasticity and linearity were also assessed in this study, but neither detected any significant deviation from the model class. Thus, the logBF comparing the embellished model, the MDM with lags 4 and 2 for Regions 1 and 4, respectively, and change points for Region 1, with the original model, RS-DAG4, was evaluated as almost 9, providing the highest evidence for the embellished model.

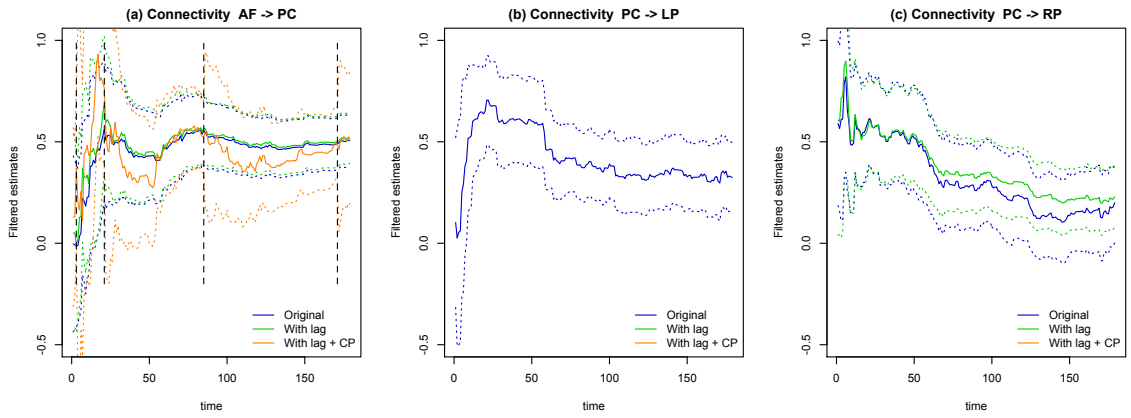


Figure 4.23: The filtered posterior mean (solid lines) with 95% HPD interval (dotted lines), considering the original MDM (blue lines), *i.e.* RS-DAG4, the MDM with lags 4 and 2 for Regions 1 and 4, respectively (green lines), and this latter model with also change points for Region 1 (orange lines), for connectivities (a) $2 \rightarrow 1$, (b) $1 \rightarrow 3$, which is the same whatever the model, and (c) $1 \rightarrow 4$, which is the same for the second and third fitted model (*i.e.* green is equal to orange lines). The vertical dashed lines represent the four change points.

In order to search networks for subjects and sessions, the scores of all possible sets of parents for every node was found per dataset. These individual scores were then summed over all datasets, and so the learning process was led. It means that we are searching network considering the same graphical structure for all subjects but with different connection strengths. The MDM-IPA provides a similar graph to RS-DAG4 (Figure 4.17, MDM-IPA), *i.e.* the information flows in a backward way, except for the edge $RP \rightarrow PC$. Therefore, this result is consistent with scientific beliefs. We also applied the MDM-DGM that allows us to model cyclic relationships. Although this has less formal validity (see discussion in Section 3.3.2), it still gives us useful heuristic information about the underlying process. But, for this data, the MDM-DGM is perhaps less useful: it appears to be oversensitive in its detection of causal interactions between the brain regions, and so provides too dense a graph where every node is connected to every other node (see Figure 4.17, MDM-DGM).

4.5 Discussion

In Section 2.3 we described some methods used to estimate the connectivity. We then compared some of these methods to the LMDM in theory, in Section 3.4, and using synthetic data, in Section 4.2. Some dynamic models, such as DCM, LDS and BDS, assume that effective connectivity is estimated by the interaction between the quasi-neural level variables (rather than the observed variables). Moreover, these models write the observed fMRI signals as a

function of the convolution matrix $\mathbf{\Phi}$ (see Section 2.3). Methods that do not consider these two features, *i.e.* the interaction between latent variables and the convolution matrix, nevertheless appear to correctly identify the effective connectivity in a synthetic dataset, which was obtained under these assumptions. We compared here different connectivity estimation approaches based on the DCM synthetic dataset, and concluded that the ones based on the LMDM were some of the most successful methods for detecting network connections, and also for estimating connection directionality. There appeared to be no significant differences in the estimation of effective connectivity when the interaction was among observation or latent variables, or when the model included the convolution matrix or not.

The MDM-IPA appears to work well when applied in a simulated model, *i.e.* when we know the truth. Especially, in the light of the analysis above, we can be fully confident that if the model class is broadly correct and with the sorts of size of dataset we are using in our experiments, our model selection should be informative.

Using real fMRI data, we demonstrated that the brain connectivities typically change over time. However, we saw (especially in Section 2.3) that most methods estimate static connections. Methods that allow the connectivity to vary over time usually estimate functional connectivity, using a sliding time window (Chang and Glover, 2010; Allen *et al.*, 2012; Leonardi *et al.*, 2013). Few methods can estimate a dynamic effective connectivity quickly and formally: they generally use approximate inferential methods, complicating the search network (Bhattacharya *et al.*, 2006; Havlicek *et al.*, 2010). Therefore, in general, no other competing class of models to the MDM is sufficiently compact, provides formal scores in a closed form, and simultaneously allows connectivity to change over time in the way the MDM can model change.

We also demonstrated here that diagnostic statistics for checking and where necessary adapting the whole class is straightforward. For instance, the feed-back intervention seems to deal adequately with identified change points (see Figure 4.22, the second column). Clearly, these diagnostics can be refined, for example by using the full power of switching state space models (see *e.g.* Fruhwirth-Schnatter, 2006, chapter 13) to model these apparent phenomena more formally. However this would also necessarily add to the complexity of the method.

The iterative modifications illustrated in the application add some additional complexity. However, they also allow us to improve the model predictions and hence refine

the analysis to allow for known phenomena such as change points appearing in the signals. Therefore, they enable us to improve the selection process without entering into the types of complex numerical estimation methods, as discussed above. In particular, improvements of the model gives us a different and higher scoring model, which is scientifically plausible and whose score can still be calculated in closed form.

Chapter 5

Group analysis using the MDM

Connectivity studies are usually based on fMRI experiments with many subjects. The use of group analysis to assess the integration of activity in brain regions brings two advantages (Mechelli *et al.*, 2002). Firstly, it enables us to investigate directly how connections differ across subjects. For instance, connection strength may vary according to age. Secondly, the degrees of freedom increase with the number of subjects. If subjects are sufficiently homogenous, this then improves the estimation process. Most methods of group analysis can be classified in four approaches, as discussed in Section 5.1. The first three approaches are *virtual-typical-subject* (VTS), *common-structure* (CS) and *individual-structure* (IS), which are developed in the context of the MDM, in Section 5.2. This, to our knowledge, has never been done before.

These methods usually assume exchangeability so that the group of subjects is a sample of the same population. Because this is not always true, the *Group-structure* approach (GS) aims to find homogeneous subgroups according to connectivity maps. In Section 5.3, we suggest a cluster analysis with a novel separation measure based on the model selection criterion, the Bayes factor (Jeffreys, 1961). We then compare these approaches using synthetic data, in Section 5.4, and real fMRI data in Section 5.5, clarifying the pros and cons of each method.

5.1 Background

The number of experiments with multiple subjects has been increasing recently. In general, four approaches that deal with multi-datasets may be found in the neuroimaging literature

(*e.g.* Mechelli *et al.*, 2002; Li *et al.*, 2008; Ramsey *et al.*, 2010; Gates and Molenaar, 2012). The first approach is the *virtual-typical-subject* (VTS) which ignores the inter-subject variability, assuming that the information from different datasets come from the same subject. This "typical subject" can be found by calculating the average of observed variables for every node over subjects or concatenating the datasets, so that methods designed for a single individual can be used (Zheng and Rajapakse, 2006; Rajapakse and Zhou, 2007; Li *et al.*, 2008). When datasets are concatenated, it increases the number of data points per node and consequently the degree of freedom is higher to estimate the parameters. However, the assumptions of this approach, *i.e.* "variations in connectivity from subject to subject are random, well-behaved and uninteresting", are not always true (Mechelli *et al.*, 2002). Moreover, the variability of *concatenated data* may be significantly higher than individual variability whilst the variability of *averaged data* may be very much lower than the usual variability found for each subject. Therefore, the results of this group-based analysis may not reflect some of the features found in the individual context (Gonçalves *et al.*, 2001). In addition, it is not possible to compare the interactions by different characteristics, such as task performance or gender.

Gates and Molenaar (2012) gave two reasons in which the VTS is not suitable for modelling a brain network. The first concerns the connectivity strength that is in general expected to vary over subjects. For instance, some researchers wish to study the relation between the *connectivity strength* and *disease level*, and this cannot be addressed using this method. Secondly, the communication pattern among brain regions may differ from individual to individual. For instance, in a study of fMRI activation pattern related to writing, three of five regions showed inconsistent results across subjects (Sugihara *et al.*, 2006). To address these problems a second method, *common-structure* (CS), and a third method, *individual-structure* (IS), have been proposed respectively.

The CS approach considers the same network structure but allows the parameters to differ between subjects. The connectivity strengths are expected to vary over subjects due to measurement error or the individual characteristic of influences from one region to another. An example of this approach is the *Independent Multiple-sample Greedy Equivalence Search* (IMaGES) which uses BIC scores to find a Markov equivalence class, basically summing the scores over subjects, considering the same graphical structure (Ramsey *et al.*, 2010). Clearly

the CS approach cannot, therefore, consider the pattern of connectivity may change over subjects. However, this may well happen, for example, in a resting-state experiment when people are free to think of anything, and someone may use their memory whilst others may do calculations. Another reason why this might be violated is when a group of patients have a disease in different degrees of severity. This can then result in different connections arising between brain regions (Li *et al.*, 2008).

The next approach, *individual-structure* (IS), drives the learning network process individually in each dataset so that results are pooled into a single network. Usually the group network is formed by edges that exist for the greatest number of subjects. An example of this approach is Oates (2013) who proposed an algorithm that firstly scored individuals, then a group network was found by minimising the distance between the individual and the group network. Although the IS approach seems to cope well with the different interactions, its results are often inconsistent among subjects if they form a heterogeneous group (Gonçalves *et al.*, 2001; Mechelli *et al.*, 2002). We show an example of this in Section 5.4.

Li *et al.* (2008) compared these three approaches using a DBN and BIC scores. They concluded that the group-level results may vary considerably among approaches. Moreover, it is not possible to say which method is generally superior over the others as long as the interpretation of results depends directly on the assumptions of each method (Li *et al.*, 2008). However, although some of these methodologies explicitly recognise the intra-subject variability, none of them assesses the homogeneity of the individual connectivity maps. If the approaches described above are applied in a heterogeneous group, then conclusions based on group network may be misleading. A fourth approach, *group-structure* (GS), has been proposed which aims to deal with a type of heterogeneity as explained below.

The GS approach studies the group homogeneity through a cluster analysis, considering a particular measure of similarity between subjects. If this analysis suggests that subjects should be clustered into disjoint subgroups, then this group of subjects is not homogeneous. When heterogeneity is present, it is suggested that any subsequent analyses of interest should be done for every subgroup independently. Note that no prior classification information is necessary to use these methods. For instance, Kherif *et al.* (2004) defined a separation measure between two subjects' data based on a multivariate correlation. Another example, Gates (2012) used the IS approach to estimate the effective connectivity of

both individual and group network, using the Group Iterative Multiple model estimation (GIMME; Gates and Molenaar, 2012). In this work, Gates proposed the correlation between connectivity strengths as the separation measure between subjects. In this thesis, we propose an alternative separation measure between subjects as a function of the model selection criterion, Bayes factor.

5.2 The VTS, the IS and the CS Applied to the MDM

The VTS approach finds a typical subject, assuming the same network with exactly the same connectivity for all subjects. Within an MDM framework, the "typical subject" was defined by first calculating the average of time series variables over all subjects. Based only on this "ordinary subject", the search method, such as the MDM-IPA or the MDM-DGM, can then be used to find the group network. Note that the local score for node r can now be written as:

$$c_a\left(r, \bar{M}(r)\right) = \sum_{t=1}^{T} \log p(\overline{y}_t(r)|\overline{\mathbf{y}}^{t-1}, \overline{\mathbf{x}}_t(r), \bar{M}(r)),$$

where $\overline{y}_t(r)$ is the average of observed variables at time t and node r over subjects, $\overline{\mathbf{y}}^t = (\overline{\mathbf{y}}_1, \ldots, \overline{\mathbf{y}}_t)'$, $\overline{\mathbf{y}}_t = (\overline{y}_t(1), \ldots, \overline{y}_t(n))'$, and $\overline{\mathbf{x}}_t(r) = (\overline{y}_t(1), \ldots, \overline{y}_t(r-1))'$. Here $\bar{M}(r)$ is the model defined by the parent set of node r so that the group network consists of $\bar{M} = (\bar{M}(1), \ldots, \bar{M}(n))$. The connectivity strength for the group network is estimated based on the smoothed posterior distribution of parameters θ's considering the MDM fitted for this typical subject.

The CS approach assumes that all subjects share the same group network structure, but the parameters may differ over subjects. In this way, the parameter estimation process is applied for each subject independently. Then, for the same model $\bar{M}$, the scores used in search process is defined as:

$$c\left(r, \bar{M}(r)\right) = \sum_{i=1}^{S} \sum_{t=1}^{T} \log p(y_{it}(r)|\mathbf{y}_i^{t-1}, \mathbf{x}_{it}(r), \bar{M}(r)), \tag{5.1}$$

where S is the number of subjects, $y_{it}(r)$ is the observed variable for region r and subject i at time t, $\mathbf{y}_i^{t-1}$ is the observed cumulative data until time $t-1$ for subject i, and $\mathbf{x}_{it}(r) =$

$(y_{it}(1), \ldots, y_{it}(r-1))'$. Note that considering the CS approach the parents of a particular node r are the same for all subjects ($\bar{M}(r)$). The group network is estimated by a search algorithm using these scores of equation (5.1). The MDM is then fitted for each subject using the same graphical structure $\bar{M}$, and the connectivity strength for group network is estimated as the average of the smoothed estimates of θ's over subjects.

The IS approach usually learns individual networks independently, using individual scores

$$c_i\left(r, M_i(r)\right) = \sum_{t=1}^{T} \log p(y_{it}(r)|\mathbf{y}_i^{t-1}, \mathbf{x}_{it}(r), M_i(r)), \tag{5.2}$$

where $M_i(r)$ is the model defined by the parent set of node r for subject i so that the individual network for subject i consists of $M_i = (M_i(1), \ldots, M_i(n))$. Then the group network structure ($\bar{M}$) consists of the edges that exist in the individual network for most subjects. The MDM is then fitted for all subjects using the group network and, as in the CS approach, the connectivity strength may be estimated as the average of the smoothed estimates of θ's over subjects.

5.3 Clustering with Pairwise Log Bayes Factor Separation

In the group-structure approach (GS) subjects are first grouped according to the similarities in their graphical structures. These similarities are defined by a separation measure, $d(i,j)$, calculated for every pair of subjects i and j, comparing the individual networks, M_i, with the pairwise group network, m_G, as follows,

$$d(i,j) = c_{ij}\left(m_I\right) - c_{ij}\left(m_G\right),$$

where $m_I = (M_i, M_j)$,

$$\begin{aligned} c_{ij}\left(m_I\right) &= \sum_{r=1}^{n} \left(c_i\left(r, M_i(r)\right) + c_j\left(r, M_j(r)\right)\right), \\ c_{ij}\left(m_G\right) &= \sum_{r=1}^{n} \left(c_i\left(r, m_G(r)\right) + c_j\left(r, m_G(r)\right)\right), \end{aligned}$$

$m_G = (m_G(1), \ldots, m_G(n))$, for $i \in \{1, \ldots, S-1\}$, $j \in \{2, \ldots, S\}$, $j > i$. Here the individual networks, M_i, are estimated by maximising the scores in equation (5.2). The pairwise group networks, m_G, is estimated by maximising the sum of scores for only two subjects, i and j, such as in equation (5.1), considering $\bar{M}(r) = m_G(r)$.

Some properties of $d(i, j)$ are given below.

1. For the MDM-IPA, the scores are exactly the LPL. Then $d(i, j)$ can be seen as the logBF comparing the model that assumes subjects i and j have different graphical structures with they share the same one. Thus, we call this separation measure as *the pairwise logBF separation.*

2. The pairwise logBF separation is symmetric, i.e. $d(i, j) = d(j, i)$.

3. If the estimated individual graphical structures for subjects i and j are the same, then $d(i, j) = 0$. As M_i is the network that maximises the scores in equation 5.2, then $c_i(M_i) > c_i(M_i^*)$, where $c_i(M_i) = \sum_{r=1}^{n} c_i(r, M_i(r))$ and M_i^* is any possible network for subject i, except M_i. Thus,

$$c_i(M_i) + c_j(M_j) > c_i(M_i^*) + c_j(M_j^*).$$

By definition, the pairwise group network assumes that both subjects share the same graphical structure, *i.e.* $M_i^* = M_j^* = m_G^*$. Thus, when $M_i = M_j$, the above inequality becomes

$$\begin{aligned} c_i(M_i) + c_j(M_i) &> c_i(M_i^*) + c_j(M_i^*) \\ c_{ij}(M_i) &> c_{ij}(m_G^*). \end{aligned} \tag{5.3}$$

As m_G^* is a network other than M_i and given equation (5.3), M_i is the pairwise group network that maximises the scores in equation 5.1, *i.e.* $m_G = M_i$. Therefore,

$$d(i, j) = \sum_{r=1}^{n} \left(c_i(r, M_i(r)) + c_j(r, M_j(r))\right) - c_{ij}(m_G),$$

as $M_i = M_j$, then $d(i, j) = c_{ij}(M_i) - c_{ij}(m_G)$,

and as $M_i = m_G$, then $d(i, j) = 0$.

4. By definition, the separation $d(i,j)$ is non-negative. This is, because $c_k(M_k) \geq c_k(M_k^*)$, whenever M_k is selected by maximising the scores in equation 5.2, and M_k^* is any possible individual network for subject k (now M_k^* can also be M_k), for $k = i, j$. Thus,

$$c_i(M_i) + c_j(M_j) \geq c_i(M_i^*) + c_j(M_j^*)$$

letting $M_i^* = M_j^* = m_G$, $c_i(M_i) + c_j(M_j) \geq c_{ij}(m_G)$

$$c_{ij}(m_I) - c_{ij}(m_G) \geq 0$$

$$d(i,j) \geq 0.$$

Using a cluster analysis with these pairwise logBF separations, subjects are grouped according to their similar networks. Then equation (5.1) is used to score models for subjects belonging to the same subgroup and so a graphical structure is estimated for each subgroup independently, *i.e.* $\bar{M}_1, \ldots, \bar{M}_G$, where G is the number of subgroups. The connectivity strength is estimated per subgroup as the average of estimated parameters θ's over subjects belonging to the same subgroup.

5.4 Comparing Methods Using Synthetic Data

In this section, we compare the four group analysis approaches described above using synthetic data. The aim of this section is to assess the efficiency of methods when subjects are sampled from populations whose individuals may exhibit different networks.

5.4.1 Simulating Data

Data were simulated from 3 different DAGs (DAG1, DAG2 and DAG3 in Figure 5.1), 10 subjects for each DAG, and considering 4 nodes and 197 time points as described below.

Firstly for all DAGs we set

$$\theta_{tid}^{(k)}(r) \sim \mathcal{N}(\theta_{t-1id}^{(k)}(r), W_d^{(k)}(r)),$$

for $r = 1, \ldots, 4$; $t = 1, \ldots, 197$; $i = 1, \ldots, 10$; $d = 1, 2, 3$; $k = 1, \ldots, p_{rd}$; and $W_d^{(k)}(r) = 0.04 \times V_d(r)$.

For DAG1, $p_{11} = 1$; $p_{21} = p_{31} = 2$ and $p_{41} = 3$. The initial values ($t = 0$) for

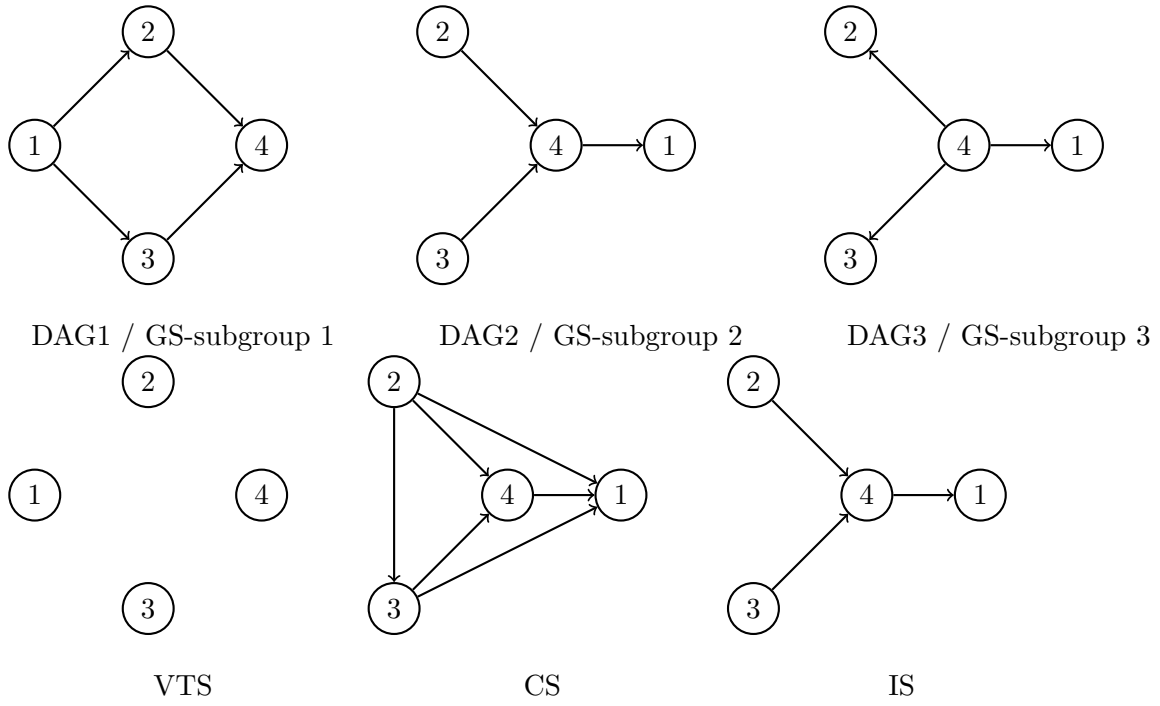


Figure 5.1: Data was simulated considering these three difference graphical structures: DAG1, DAG2 and DAG3 (in the first row). Considering these synthetic data, four methods were used to estimate network: the VTS, the CS, the IS and the GS approach. The GS approach found three groups in which their estimated graphs coincide with true DAGs, as shown in the first row. The estimated DAGs for other approaches are in the second row.

the regression parameters were 0.2 for connections $Y(1) \rightarrow Y(2)$ and $Y(3) \rightarrow Y(4)$; 0.4 for $Y(1) \rightarrow Y(3)$; 0.3 for $Y(2) \rightarrow Y(4)$ and the value 0 for other θ's (intercept parameters). The observational variance ($V_1(r)$) was defined as almost one for all nodes. Observed values were then simulated using the following equations:

$$\begin{aligned} Y_{tid}(1) &= \theta_{tid}^{(1)}(1) + v_{tid}(1); \\ Y_{tid}(r) &= \theta_{tid}^{(1)}(r) + \theta_{tid}^{(2)}(r) Y_{tid}(1) + v_{tid}(r), \qquad r = 2, 3; \\ Y_{tid}(4) &= \theta_{tid}^{(1)}(4) + \theta_{tid}^{(2)}(4) Y_{tid}(2) + \theta_{tid}^{(2)}(4) Y_{tid}(3) + v_{tid}(4); \end{aligned}$$

where $d = 1$ and $v_{tid}(r) \sim \mathcal{N}(0, V_d(r))$, for $r = 1, \ldots, 4$.

For DAG2, $p_{12} = 2$; $p_{22} = p_{32} = 1$ and $p_{42} = 3$. The initial values were 0.3 for connection $Y(2) \rightarrow Y(4)$; 0.2 for $Y(3) \rightarrow Y(4)$; 0.4 for $Y(4) \rightarrow Y(1)$ and again the value 0 for other θ's (intercept parameters). The observational variance ($V_2(r)$) was also defined as

almost one for all nodes. Observed values were then simulated using the following equations:

$$\begin{aligned} Y_{tid}(r) &= \theta^{(1)}_{tid}(r) + v_{tid}(r), \qquad r = 2, 3; \\ Y_{tid}(4) &= \theta^{(1)}_{tid}(4) + \theta^{(2)}_{tid}(4) Y_{tid}(2) + \theta^{(2)}_{tid}(4) Y_{tid}(3) + v_{tid}(4); \\ Y_{tid}(1) &= \theta^{(1)}_{tid}(1) + \theta^{(2)}_{tid}(1) Y_{tid}(4) + v_{tid}(1); \end{aligned}$$

where $d = 2$ and $v_{tid}(r)$ is defined as before.

For DAG3, $p_{13} = p_{23} = p_{33} = 2$ and $p_{43} = 1$. The initial values were 0.6 for connection $Y(4) \to Y(1)$; 0.5 for $Y(4) \to Y(2)$; 0.2 for $Y(4) \to Y(3)$ and intercept parameters receive the value zero. The observational variance ($V_3(r)$) was almost 0.3 for all nodes. Observed values were then simulated using the following equations:

$$\begin{aligned} Y_{tid}(4) &= \theta^{(1)}_{tid}(4) + v_{tid}(4); \\ Y_{tid}(r) &= \theta^{(1)}_{tid}(r) + \theta^{(2)}_{tid}(r) Y_{tid}(4) + v_{tid}(r), \qquad r = 1, 2, 3; \end{aligned}$$

where $d = 3$ and $v_{tid}(r)$ is defined as before.

5.4.2 The GS Approach

The pairwise logBF separation for all pair of subjects was evaluated as shown in Section 5.3 and considering the MDM-IPA. Then, to assess the homogeneity of this group, we used *hierarchical cluster* and *multidimensional scaling* (MDS), see below. The hierarchical agglomerative clustering method provides successive nested clusters rather than a particular partition (Everitt *et al.*, 2011, Chapter 4). This allows the series of partitions to range between the first stage with S clusters (one per subject) and the last stage with a single cluster (containing all subjects). Here we are using the well-known *complete linkage* (Everitt *et al.*, 2011, Chapter 4). The first stage consists of S clusters and so, subjects with the smallest pairwise separation between them form the same cluster, in the second stage. From the third stage, the separation between two clusters (with one or more subjects) is evaluated as the maximum of all pairwise separation between subjects, where pairs consist of one subject from each cluster. Then two clusters that have the smallest separation, comparing to other pairs of clusters, are joined to form a new cluster. This procedure continues until all subjects are included in the same cluster (Everitt *et al.*, 2011, Chapter 4).

The hierarchical classification results can be illustrated through a *dendrogram* which shows the process and the partitions found at each stage. To define subgroups, we are using the *dynamic tree cut* (hybrid algorithm) which works in two steps (Langfelder *et al.*, 2008). In the first step, this algorithm builds some clusters considering the information from the dendrogram (*e.g.* height where a particular cluster joins to all other clusters). In the second step, the unassigned subjects (*i.e.* subjects who were not previously included in any cluster in the first step) are assessed as to whether they may belong to some cluster using only the separation measures. Subjects who continue to be unassigned in the second step are considered as outliers (see details about the dynamic hybrid algorithm in Appendix C).

Figure 5.2 (left) shows the dendrogram which was found using the R packages *hclust* and *dynamicTreeCut*, and the minimum size of cluster equal to 3. In this diagram, subjects are represented by the number of their respective DAG. The hybrid algorithm identifies correctly the number of subgroups, *i.e.* the three coloured rectangles under the dendrogram. Most of the subjects were correctly grouped (only three subjects — numbered with asterisks — are in the wrong group). The criteria of logBF ≥ 2 shows a strong evidence for the first model used in its calculation (West and Harrison, 1997). Recall that, in the calculation of $d(i, j)$, the first model is that individual DAGs were estimated independently. Therefore, because the average separation between subjects belonging to the same group was around 1.8, this result indicates that these subjects are likely to share the same network structure. In contrast, the average separation between groups was almost 30 and so this shows a strong evidence for people from different subgroups having different graphical structure.

Next multidimensional scaling (MDS) was explored in this context. The MDS depicts patterns in the separation between subjects by a visual representation (see details in Appendix C). By using geometry, the best separation between subjects in a low-dimensional scaling is used to represent the original dissimilarity measure (Everitt *et al.*, 2011, Section 2.3.3). Note that, in the MDS plot, it is possible to recognise subgroups and outliers, and also to verify a measure of the quality of this approximate Euclidian depiction. For instance, Figure 5.2 (right) shows a 2D plot which captured almost 80% of this information, where subjects are labelled by the number of their DAG as before. Clearly subjects from DAG1 are on the right of the figure, whilst subjects from DAG2 are on left and DAG3 in the centre, although note that some subjects from DAG3 and DAG2 are close to each other in the graph.

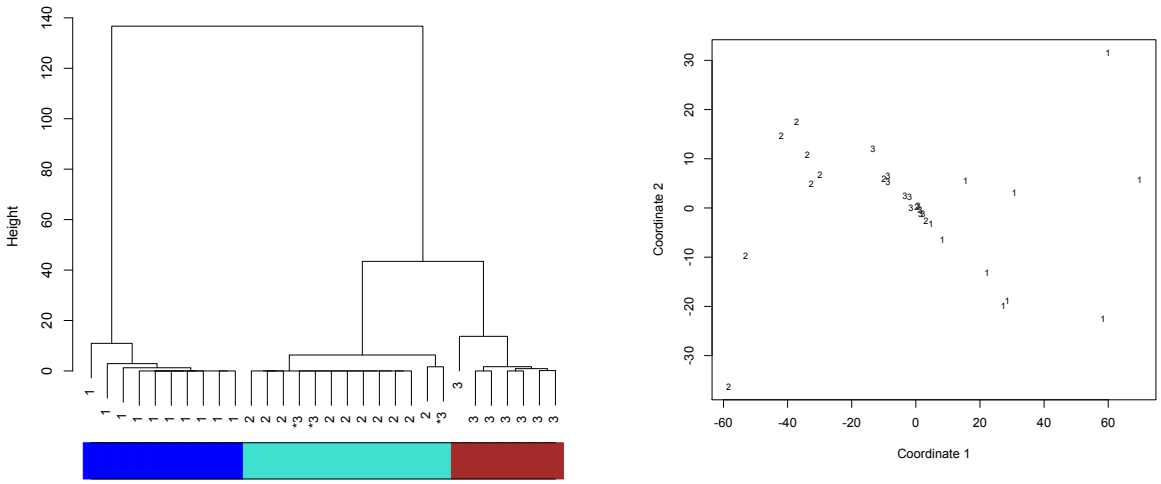


Figure 5.2: Dendrogram (*left*) and MDS (*right*) for synthetic dada using the pairwise logBF separation. Numbers 1, 2 and 3 correspond to subjects simulated based on DAG1, DAG2 and DAG3 respectively. Coloured rectangles under the dendrogram identify the three subgroups found by the hybrid algorithm. Subjects marked with asterisks were included in a subgroup containing most of subjects generated from a different model.

5.4.3 Comparing Group Analysis Approaches

The graphical structures for VTS, CS and IS approaches were estimated as described in Section 5.2, considering the MDM-IPA (see Figure 5.1). Surprisingly there was no connection for the VTS approach — the average of time series over subjects nullified connections in some way. In contrast, the result for the IS approach gave no surprises whenever it picked up the edges that exist for most individuals. Indeed this result showed exactly the three connections that exist for two-thirds of subjects. In this example, for a heterogeneous group, the CS approach overestimated the number of edges, *i.e.* with the three most popular connections plus three erroneous connections.

In contrast to other methods, the GS approach identified correctly different networks. Three subgroups were formed by 10, 11 and 9 subjects in subgroup 1, 2 and 3, respectively (Figure 5.2). As a result the estimated graphical structure is the same as the true DAG for every group, as shown in Figure 5.1.

Considering connectivity strength estimates, the results also indicated the GS approach as the most effective method for this study. Figure 5.3 provides the true value of the regression parameter and the average of smoothed posterior mean over subjects. For instance, considering the connectivity $Y(3) \rightarrow Y(4)$ at around the time 120, the true connection of subgroup 1 was almost 0 whilst the true value of subgroup 2 was -2. There was

no large difference between true and estimated values for the GS approach, subgroups 1 and 2. However, the estimated value was around -1 for the CS and the IS approaches, which was indeed the average between the true values of subgroup 1 and subgroup 2. In general, the GS approach provided estimates closer to true values than other methods.

5.4.4 Comparing Separation Measures

We used four other measures of a graphical separation and compared these with the pairwise logBF separation. The first is *indegree* which is defined as the number of edges that arrives in a node, *i.e.* the number of parents. In contrast *outdegree*, the second measure, is the number of edges that leaves from a node, *i.e.* the number of children. To find the third measure, *degree*, we firstly transformed the estimated DAG into an undirected graph. We then calculated the number of edges connected to a particular node, and the separation between two subjects was found by Euclidian distance. The fourth measure is the Structural Hamming Distance (SHD) defined as the number of directed edges that exist in one graph but not in the other plus the number of edges that exist in the latter graph but not in the former.

Figure 5.4 shows the dendrogram for these 4 new separations. The hybrid algorithm identified correctly three subgroups without outliers (represented by grey rectangles) only for outdegree separation. The outdegree separation showed the best performance amongst these four separations, but performed slightly worse than the pairwise logBF separation, with 4 misclassified subjects. The worst result was found for indegree separation where 10 out of 30 subjects were misclassified. It is important to highlight that the pairwise logBF separation has the advantage of being based on the well-known model selection measure and is, therefore, simpler to interpret at least from a Bayesian perspective.

5.5 Group-structure using the Real RS fMRI Data

Here we are considering again the resting-state study described in Section 4.4. Recall that these real fMRI data consist of 197 fMRI resting-state time-points and 4 ROI's: Posterior Cingulate - PC; Anterior Frontal - AF; Left Lateral Parietal - LP and Right Lateral Parietal - RP. Information is available for 3 sessions for each one of 25 subjects, acquired under the same experimental conditions. As shown in Section 4.4, firstly four different graphical

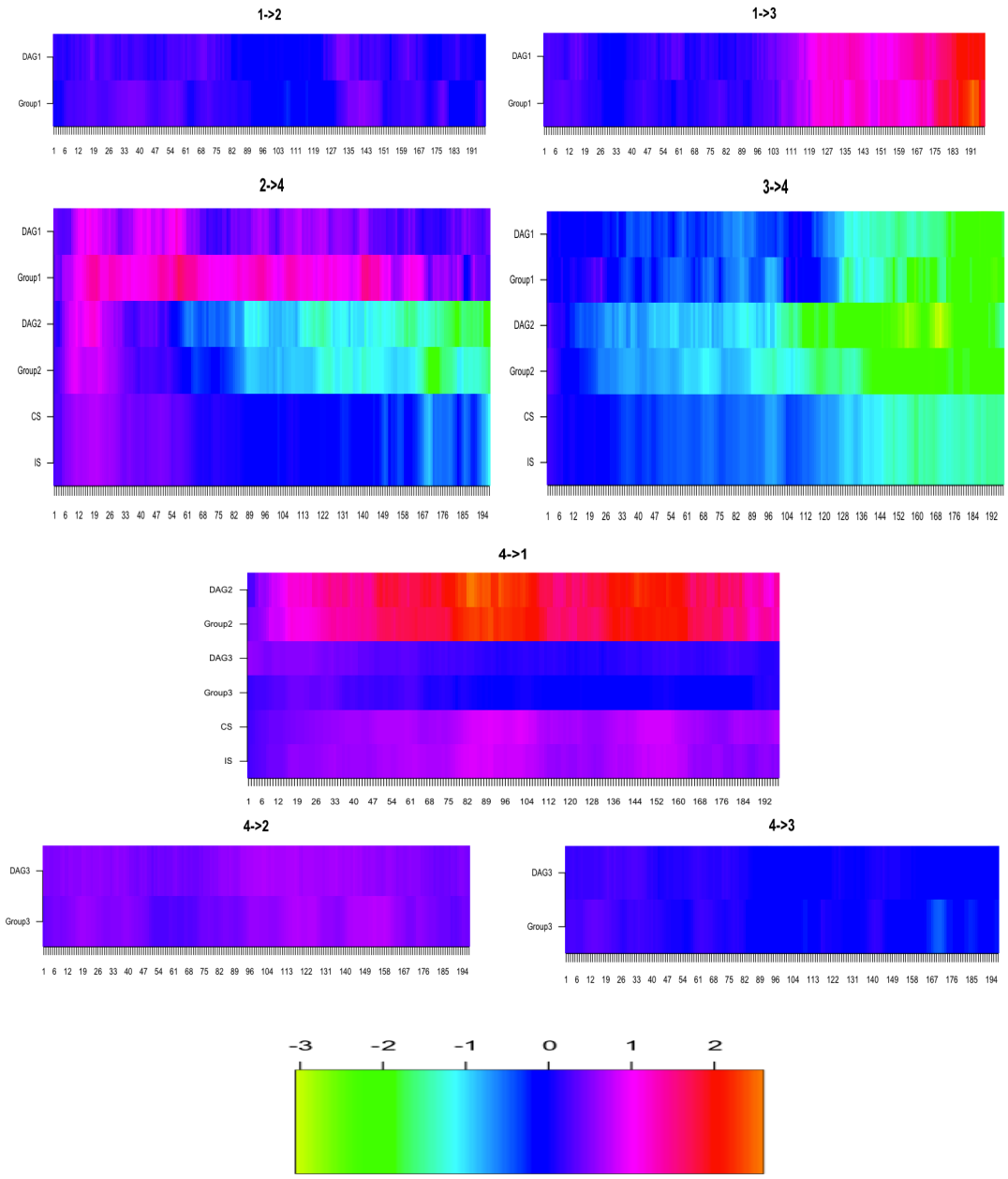


Figure 5.3: DAG1, DAG2 and DAG3 are true values of connectivity over time whilst the CS, the IS and the GS (Goup1, Group2 and Group3) are the average of smoothed posterior mean of connectivity over subjects considering the respective approach.

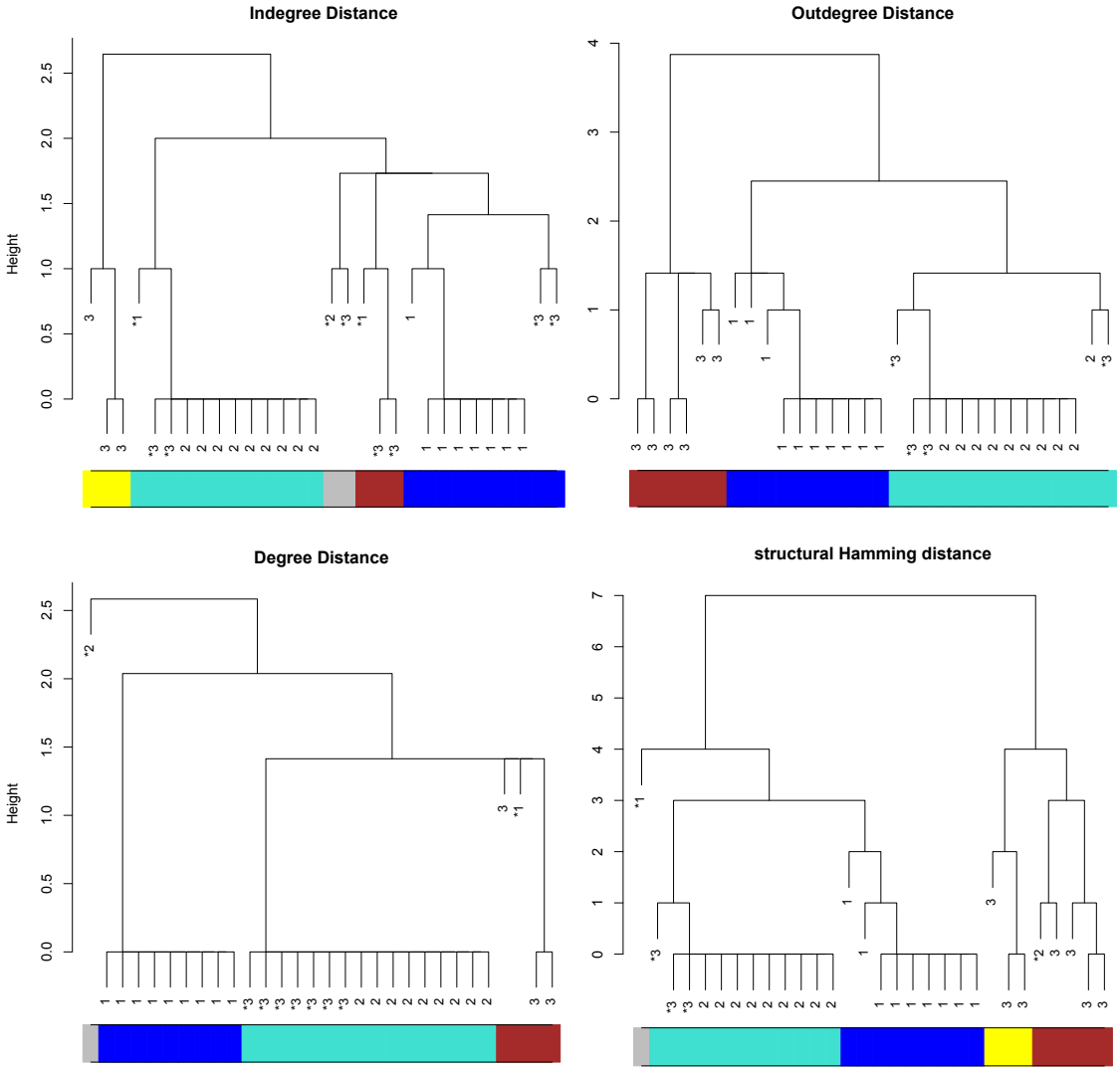


Figure 5.4: Dendrogram of synthetic dada using four separations: Indegree, Outdegree, Degree and SHD. Numbers 1, 2 and 3 correspond to subjects simulated based on DAG1, DAG2 and DAG3 respectively. Coloured rectangles under dendrogram identify subgroups found by hybrid algorithm, being outliers represented by grey colour. Subjects marked with asterisks were included in a subgroup containing most of subjects generated from a different model.

structures were chosen for representing the scientific beliefs about the brain connectivities (RS-DAG1 to RS-DAG4 in Figure 5.5). Comparing these 4 DAGs, the graphical structure that maximises the log predictive likelihood was selected for each dataset. The RS-DAG4 was chosen for most of datasets (almost 55%), following by RS-DAG1 for about 40%. Then a group analysis was applied without giving preference to any specified model. Broadly the results of the search process were consistent with scientific knowledge, as shown below.

5.5.1 VTS, CS and IS Approaches

The estimated graphical structures for VTS, CS and IS approaches are given in Figure 5.5, considering the 75 datasets and the MDM-IPA. As expected, the result of IS approach was RS-DAG4 that was the graph chosen for most subjects. In contrast to the simulation study, the VTS showed a plausible result which was close to RS-DAG4. The CS approach also provided a consistent result with the first learning network process, *i.e.* the information in these brain regions flows in a backward way. However, the directionality of the connection between RP and PC regions was contrary to what was expected. Note that none of the methods identified two different graphical structures in this population.

5.5.2 Comparing Sessions

Here for simplicity we are assuming that all sessions share the same graphical structure. However it is important to assess the reproducibility of results before proceeding with the analysis (McGonigle *et al.*, 2000; Kherif *et al.*, 2004). In Section 4.4, we have shown that the results of sessions are consistent for the same subject, *i.e.*, on average, 91% of the sessions of the same subject have the same result when the 4 original DAGs were compared (from RS-DAG1 to RS-DAG4 in Figure 5.5). Now we selected randomly four subjects and applied the GS approach. Considering the minimum size of cluster equal to 3, the hybrid algorithm suggested that there are two subgroups, where each subgroup is formed by sessions of the same subjects (Figure 5.6). Sessions of subjects *2* and *4* appeared to belong to the first subgroup whilst subjects *1* and *3* belonged to the second one, except for one of the sessions of subject *4*. Therefore, in general, sessions of the same subjects appeared close to each other using this method.

Here we showed the application of our methodology, using the pairwise separation

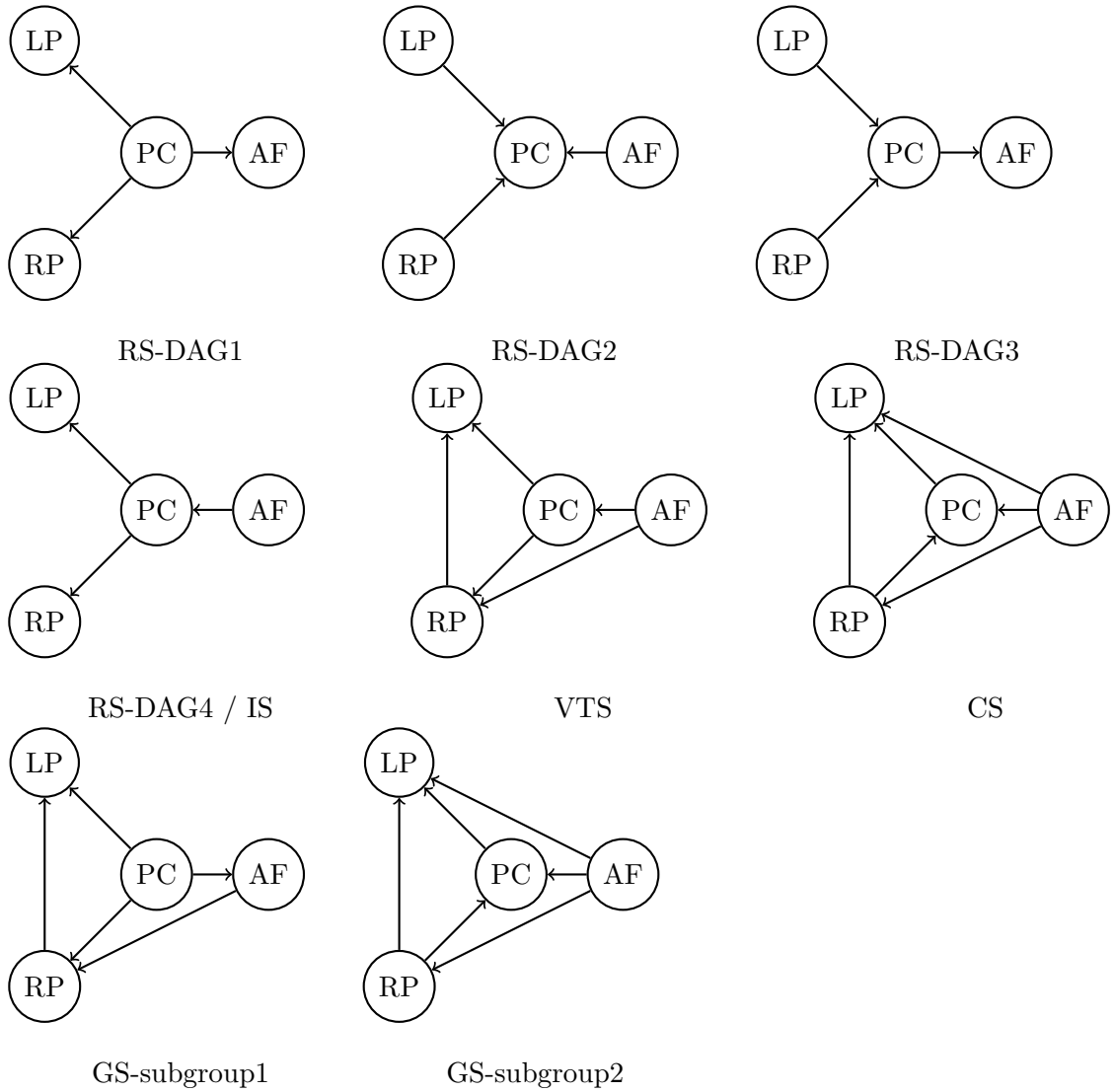


Figure 5.5: The graphical structures from RS-DAG1 to RS-DAG4 were used in the first learning process for resting state fMRI real data. RS-DAG4 was chosen for most of the datasets, around 55%, following by RS-DAG1 with almost 40%. Then the second learning process using the MDM-IPA was applied. As a result, IS approach provided the same graph as RS-DAG4, but VTS and CS also gave similar results to the most popular graph. Only GS approach shows that this group of subjects has two different graphical structures. The estimated graph of GS-subgroup1 is similar to RS-DAG4 whilst GS-subgroup2 is similar to RS-DAG1. Node PC means the posterior cingulate area, AF means the anterior frontal area, LP means the left lateral parietal area and RP means the right lateral parietal area.

measure, to verify the assumption of sessions are exchangeable. If this assumption appears to hold, as in this example above, the data from sessions can improve both the learning network process and the estimation of connectivities, for increasing the amount of information used in the analysis. However, even when the graphical structures of sessions are not expected to be exactly the same, this information can also help the group analysis as shown in Section 6.4.2.

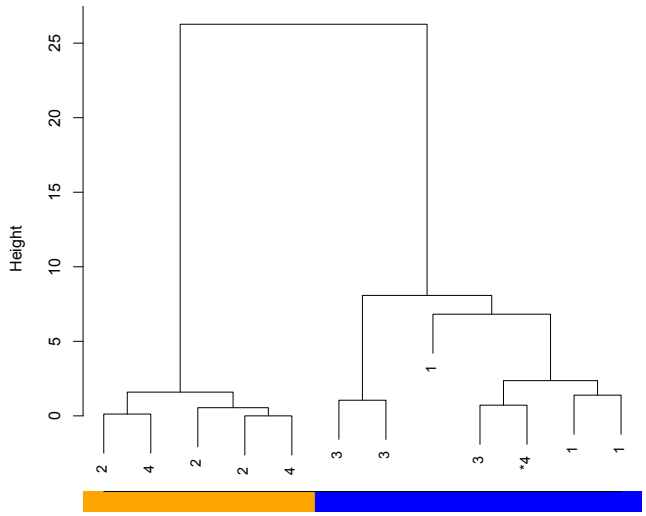


Figure 5.6: Dendogram of real fMRI dada using the pairwise logBF separation for 3 sessions of each 4 subjects selected randomly. The number corresponds to subjects and colours orange and blue correspond the two subgroups suggested by the hybrid algorithm. In general sessions of the same subjects are close each other, except for one session of subject 4 marked with asterisks.

5.5.3 The Application of the Group-structure Approach

The GS approach was applied for 25 subjects, summing the scores over sessions. Figure 5.7 (*left*) gives the result of this analysis through a dendrogram and the MDS plot, considering the MDM-IPA. The hybrid algorithm suggested two subgroups (orange and blue). The scores of subjects who belong to the same subgroup were summed and then the MDM-IPA was applied for each subgroup independently. The graphical structures are shown in Figure 5.5, GS-subgroup1 (orange group) and GS-subgroup2 (blue group). Note that the estimated graph of GS-subgroup2 is similar to RS-DAG4 whilst GS-subgroup1 is similar to RS-DAG1. Therefore, in contrast to other methods, the result of the GS approach was consistent with the previous analysis that showed evidence of two different subgroup networks. Figure 5.8 shows the average of connectivity smoothed estimate over subjects. Note that the connectivity strength also differed between subgroups. For instance the connectivity PC $\rightarrow$ LP was

stronger in subgroup 1 (second row) than in subgroup 2 (ninth row). The logBF comparing heterogeneous with homogenous group was found to be around 118, showing a strong evidence for a model where subjects were clustered into two subgroups.

5.5.4 The GS Approach with the MDM-DGM algorithm

The MDM-DGM was also applied to this resting-state data, finding parents that maximise the scores of a particular node independently of the other nodes. Figure 5.7 (*right*) shows dendrogram and MDS plot for this search algorithm. The hybrid algorithm also suggested two subgroups. All subjects of the subgroup 1 of the MDM-DGM (orange group in the right) appeared in the subgroup 2 of the MDM-IPA (blue group in the left), except for subjects 2 and 23. Indeed the estimated graphical structures of the MDM-DGM were the same as of the MDM-IPA, replacing the directed edges by bi-directed edges. That is, the subgroup 1 of the MDM-DGM (orange) had a dense graph, where all nodes were connected with all nodes (comparing with Figure 5.5 GS-subgroup2); and the subgroup 2 of the MDM-DGM (blue) had also a dense graph, except that edges between regions LP and AF did not exist (comparing with Figure 5.5 GS-subgroup1).

5.6 Discussion

Many experimental designs in neuroscience involve data collected on multiple subjects. They may differ with respect to neural connectivity, such that corresponding graphs M_i may be subject-specific (Sugihara *et al.*, 2006; Li *et al.*, 2008). Given that elements of neural architecture are largely conserved between subjects, it is natural to leverage this similarity in order to improve statistical efficiency, by addressing both the robustness of inferred graphical structure and reducing small sample bias (Mechelli *et al.*, 2002). The statistical challenge of estimating multiple related graphical models has recently received much attention, *e.g.* VTS, CS and IS approaches (Mechelli *et al.*, 2002; Zheng and Rajapakse, 2006; Rajapakse and Zhou, 2007; Li *et al.*, 2008; Ramsey *et al.*, 2010). Here the VTS was applied to evaluate the average of time series variables over subjects. Its result was poor in the synthetic study, but it was consistent with other methods for real fMRI data. The CS approach provided dense graphs, in both synthetic and real studies, *i.e.* all nodes are connected with each other. However, in the sub-groups of simulated data, where most of the subjects share the same

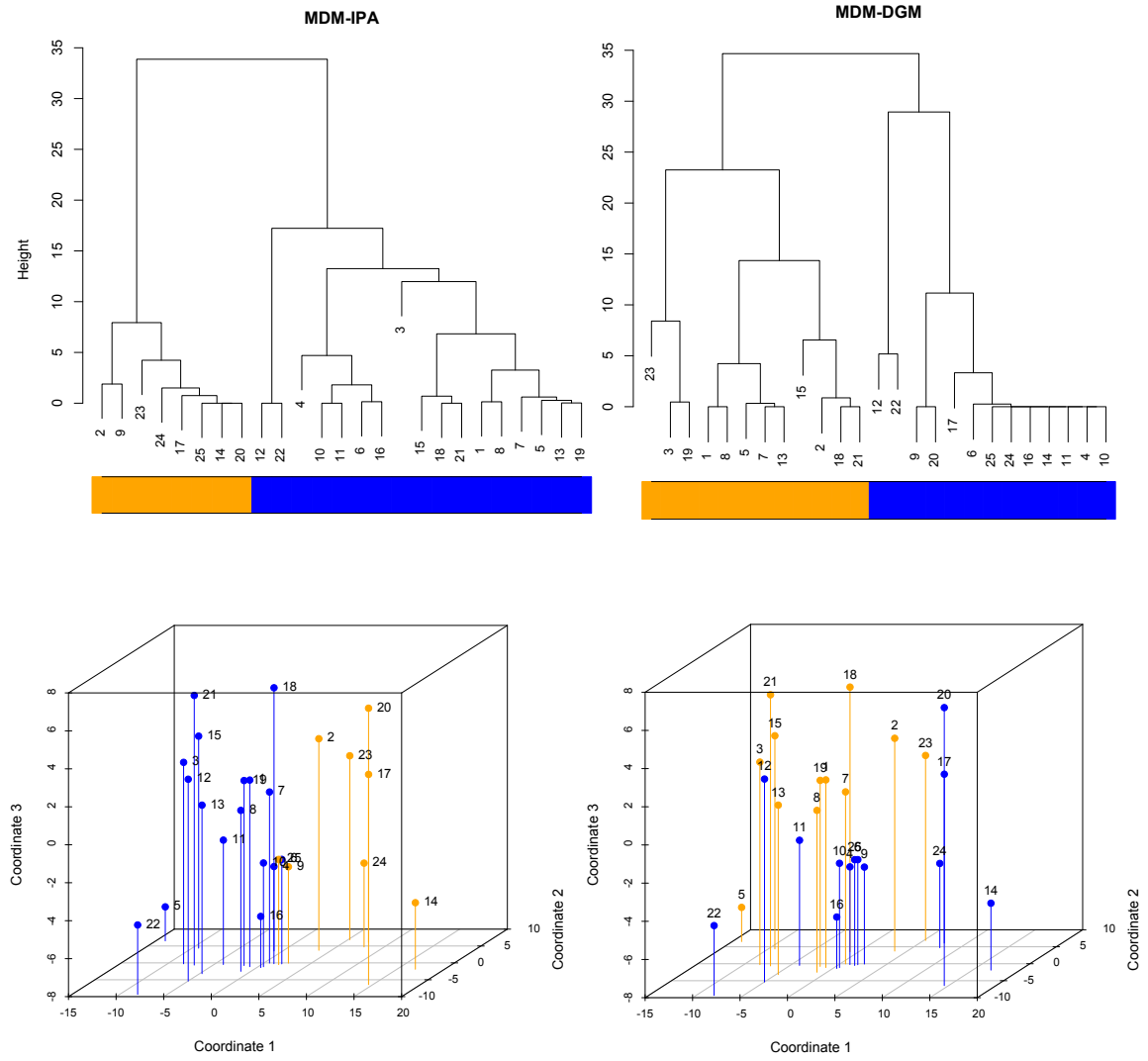


Figure 5.7: Dendrogram (*above*) and MDS (*below*) of real fMRI dada using the pairwise logBF separation for the MDM-IPA (*left*) and the MDM-DGM (*right*). Coloured rectangles under dendrogram identify subgroups found by hybrid algorithm. The MDS graph illustrates the subjects with respective colours, and captured almost 80% of the information provided by dissimilarity measure for both search algorithms, MDM-IPA (*left*) and MDM-DGM (*right*).

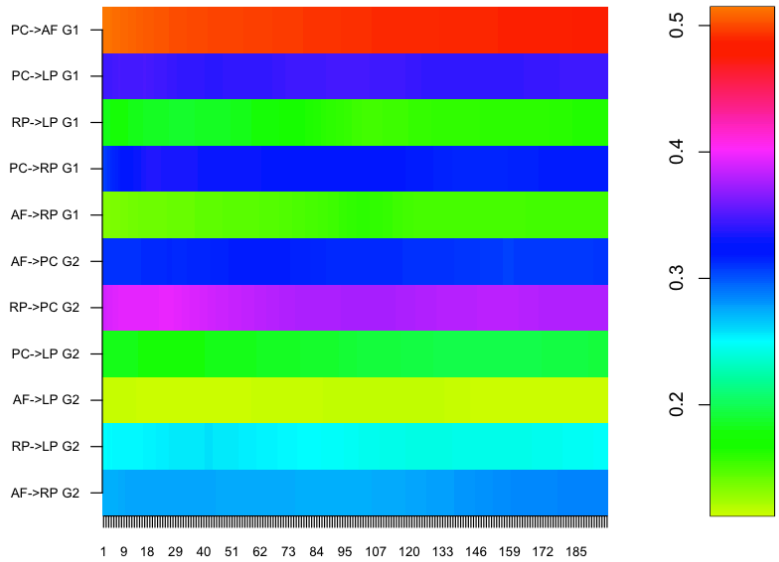


Figure 5.8: The average of connectivity smoothed estimate over subjects using real fMRI data and GS approach. G1 and G2 mean subgroup 1 and subgroup 2, using the network structure given by Figure 5.5, GS-subgroup1 and GS-subgroup2, respectively.

graphical structure, the process of summing scores had an excellent performance. The IS approach provided sparser graphs than other methods, and its result appeared to reflect what happened to most subjects.

Therefore, these studies suggest that it might be possible to increase statistical efficiency, often considerably, by formulating an appropriate joint model that couples together multiple graphs. However, these do use an exchangeability assumption, asserting the entire group is homogenous. Therefore obviously these approaches may provide inconsistent results for a heterogeneous group, as shown above. In our study, we saw that only the GS approach can recognise the heterogeneity that existed in the group. A cluster analysis using the novel pairwise logBF separation performed best when compared with other measures, such as the SHD.

Chapter 6

Estimation of multiple networks

6.1 Introduction

In the previous chapter, we discussed four approaches used to estimate a group network, considering the different ways of combining individual information. The virtual-typical-subject (VTS) approach assumes that everyone has the same graph, whilst the common-structure (CS) approach considers the same graphical structure but with connectivity strengths varying over subjects. The individual-structure (IS) approach firstly drives the learning network process for each subject independently, and then pool the results into a single network. In contrast to these methods, the group-structure (GS) approach assesses whether the group is homogeneous.

In this chapter, we discuss about how to infer both an appropriate individual and a group network. Besides the problem of heterogeneous group, we also talk about the two substantive barriers to the inference of graphical models, as described below.

Firstly, because the high variance of graphical estimators themselves, an inferred graphical structure is often not robust to reasonable perturbation of the underlying data (Claassen and Heskes, 2012). In addition, when the graphical structure is unknown, the learning algorithm adds more uncertainty to the inferential process.

Secondly, conventional model selection criteria for graphical models are often biased towards selecting more complex models (*i.e.* more edges). One reason for this is that when one model is nested in another, the larger containing model will fit the data well even if it is generated by the smaller contained model (Consonni and La Rocca, 2010), as discussed

in Section 4.2.2. Consequently many more data are required to exclude more complex alternatives. Taken together, these factors limit the extent to which neural connectivity can be accurately recovered from data.

Therefore, we present here three methods for estimating individual and group networks: the *Individual Estimation of Multiple Networks* (IEMN), the *Marginal Estimation of Multiple Networks* (MEMN), and the *Joint Estimation of Multiple Networks* (JEMN). All these methods incorporates the GS approach to deal with heterogeneous group, but the second and third methods, the MEMN and the JEMN, appears to be more robust than the IEMN, because they estimate the individual network using the information of other subjects. Moreover, the JEMN also uses a penalty function in order to provide sparser graphs. We give more details about these methods below.

In Section 6.2, we define the first method, the IEMN in which subjects are first grouped using a cluster analysis, as in the GS approach. Then, the IEMN first estimates the individual networks independently, and after that, the subgroup and the group networks are estimated using the CS approach. It therefore addresses mainly the first challenge: the heterogeneous population problem.

The second method, the MEMN, is then developed based on a method suggested by Oates (2013), in which the subgroup and the group networks are estimated using the IS approach and a similarity measure between individual and subgroup/group network structures. This method then estimates the individual networks using the information of other subjects belonging to the same homogenous subgroup. The IEMN and the MEMN are compared using real fMRI data in Section 6.3.

Following this, in Section 6.4, we present the third method, the JEMN, developed by Oates *et al.* (2014), which estimates all networks: individual, subgroup and group, at the same time. The JEMN method penalises dense graphs, addressing all challenges cited above. For the first time, we provide an application of the JEMN to real data, clarifying some of the properties of this method. Finally, we discuss the main results found in this chapter in Section 6.5.

6.2 The IEMN and the MEMN

In this section we describe a new approach for searching over MDMs which does not only estimate group networks but also individual networks, taking into account the information from other subjects. This approach is called the Marginal Estimation of Multiple Networks (MEMN) and is originally developed for DBNs, using a penalty function that represents the distance between group and individual networks (Oates, 2013). We generalise this methodology considering the GS approach, *i.e.* the cluster analysis shown above, including one more step to estimate the subgroup networks. Moreover, Oates (2013) considered the probability of a particular edge existing in his method. Here, in contrast, we develop the MEMN based on the density of a DAG. Also, Oates (2013) assumed that the parameter of the penalty function, λ, was known, being defined by scientists. It may not be easy to suggest appropriate values for this parameter, especially when the study consists of a novel experiment, and a misspecification of this parameter will provide erroneous results. To address these difficulties, we discuss here some new possibilities for estimating λ from data.

A comparison between the MEMN and the Individual Estimation of Multiple Networks (IEMN) is also provided in this section. The IEMN is basically the GS approach described in the last section, where the individual networks were estimated independently whilst the subgroup networks were estimated summing the scores over subjects belonging to the same subgroup.

Reviewing the notation, M_i is a graphical structure within the space $\mathbb{M}_i$ for subject i; $\bar{M}_g$ is a graphical structure within the space $\bar{\mathbb{M}}_g$ of subgroup $g = 1, \ldots, G$. $\mathbb{S}_g$ is the set of the indexes of subjects who belong to the subgroup g, according to cluster analysis, and S_g is the number of subjects in the subgroup g. $\bar{M}$ is a graphical structure of the group considering all subjects within the space $\bar{\mathbb{M}}$ (see Figure 6.1).

6.2.1 The Individual Estimation of Multiple Networks (IEMN)

IEMN: Individual Graphical Structures: $M_1, \ldots, M_S$

The maximum a posteriori probability (MAP) estimator of M_i considering the IEMN

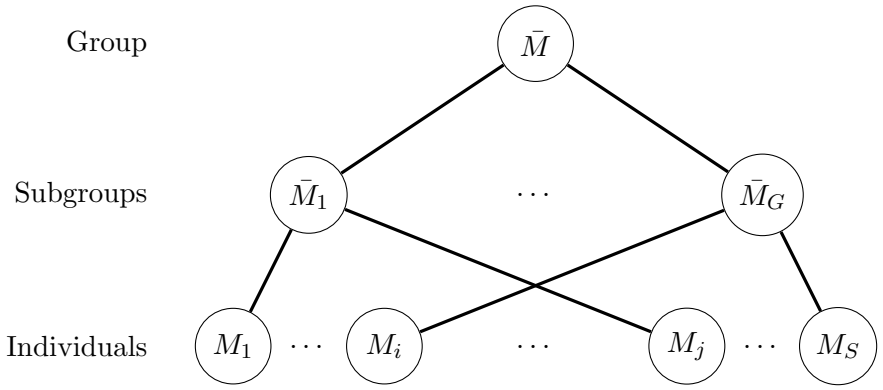


Figure 6.1: Individual networks: $M_1, \ldots, M_S$; Subgroup networks: $\bar{M}_1, \ldots, \bar{M}_G$, found by cluster analysis; and the group network: $\bar{M}$.

can be defined as

$$\hat{M}_i := \underset{M_i \in \mathbb{M}_i}{\arg\max}\, p(\mathbf{y}_i^T | M_i),$$

where $\log p(\mathbf{y}_i^T | M_i) = \sum_{r=1}^{n} c_i(r, M_i(r))$, and the score $c_i(r, M_i(r))$ is found per subject i and node r as in equation (5.2). Therefore the MDM-IPA or the MDM-DGM can be applied to find $\hat{M}_i$, per subject independently, using the scores $c_i(r, M_i(r))$.

IEMN: Subgroup Graphical Structures: $\bar{M}_1, \ldots, \bar{M}_G$

Now the MAP estimator of the subgroup network $\bar{M}_g$ is:

$$\begin{aligned}
\hat{\bar{M}}_g &:= \underset{\bar{M}_g \in \bar{\mathbb{M}}_g}{\arg\max} \prod_{i \in \mathbb{S}_g} p(\mathbf{y}_i^T | \bar{M}_g), \text{ or equivalently} \\
&:= \underset{\bar{M}_g \in \bar{\mathbb{M}}_g}{\arg\max} \sum_{i \in \mathbb{S}_g} \sum_{r=1}^{n} \sum_{t=1}^{T} \log p(y_{it}(r) | \mathbf{y}_i^{t-1}, \mathbf{x}_{it}(r), \bar{M}_g(r)) \\
&= \underset{\bar{M}_g \in \bar{\mathbb{M}}_g}{\arg\max} \sum_{r=1}^{n} c(r, \bar{M}_g).
\end{aligned}$$

So using the scores $c(r, \bar{M}_g)$, $\bar{M}_g$ can be estimated by the MDM-IPA or the MDM-DGM per subgroup independently.

IEMN: Group Graphical Structure: $\bar{M}$

For search the group network, the CS approach is applied so that the scores are summed over all subjects, as in equation (5.1). Then, the MAP estimator of the group

network considering all subjects can be defined as

$$\hat{\bar{M}} := \arg\max_{\bar{M}\in\bar{\mathbb{M}}} p(\mathbf{y}|\bar{M}),$$

where $\log p(\mathbf{y}|\bar{M}) = \sum_{r=1}^{n} c(r, \bar{M}(r))$, and $\mathbf{y}' = \left(\mathbf{y}_1^{T'}, \ldots, \mathbf{y}_S^{T'}\right)$. Again $\bar{M}$ can be estimated by the MDM-IPA or the MDM-DGM.

6.2.2 The Marginal Estimation of Multiple Networks (MEMN)

MEMN: Individual Graphical Structures: $M_1, \ldots, M_S$

This method scores individual networks based on the information from the subgroup network and the individual networks of other subjects who belong to the same subgroup, as follows.

$$\begin{aligned}
p(M_i(r)|\mathbf{y}) &= \sum_{\bar{M}_g(r)\in\bar{\mathbb{M}}_g(r)} \sum_{\substack{M_k(r)\in\mathbb{M}_k(r)\\ k\in\mathbb{S}_g\setminus\{i\}}} p(\bar{M}_g(r), M_{1^*}(r), \ldots, M_{S_g^*}(r)|\mathbf{y}) \\
&\propto \sum_{\bar{M}_g(r)\in\bar{\mathbb{M}}_g(r)} \sum_{\substack{M_k(r)\in\mathbb{M}_k(r)\\ k\in\mathbb{S}_g\setminus\{i\}}} \Big[p(\mathbf{y}|M_{1^*}(r), \ldots, M_{S_g^*}(r)) \times \\
&\quad p(M_{1^*}(r), \ldots, M_{S_g^*}(r)|\bar{M}_g(r)) \times p(\bar{M}_g(r)) \Big] \\
&\propto \sum_{\bar{M}_g(r)\in\bar{\mathbb{M}}_g(r)} \sum_{\substack{M_k(r)\in\mathbb{M}_k(r)\\ k\in\mathbb{S}_g\setminus\{i\}}} \prod_{l\in\mathbb{S}_g} \Big[p(\mathbf{y}_l^T|M_l(r)) \times p(M_l(r)|\bar{M}_g(r)) \Big] \\
&= \sum_{\bar{M}_g(r)\in\bar{\mathbb{M}}_g(r)} \Big[p(\mathbf{y}_i^T|M_i(r)) \times p(M_i(r)|\bar{M}_g(r)) \times \\
&\quad \prod_{k\in\mathbb{S}_g\setminus\{i\}} \sum_{M_k(r)\in\mathbb{M}_k(r)} \big(p(\mathbf{y}_k^T|M_k(r)) \times p(M_k(r)|\bar{M}_g(r)) \big) \Big], \qquad (6.1)
\end{aligned}$$

where a subject i belongs to the subgroup g; $k = 1^*, \ldots, S_g^*$ whenever the kth element of $\mathbb{S}_g$; and $\mathbb{S}_g \setminus \{i\}$ is the set of the indexes of all subjects belonging to the subgroup g, except for subject i. Here $\bar{\mathbb{M}}_g = (\bar{\mathbb{M}}_g(1), \ldots, \bar{\mathbb{M}}_g(n))$; and $\mathbb{M}_i = (\mathbb{M}_i(1), \ldots, \mathbb{M}_i(n))$. Note that the term $p(\mathbf{y}_i^T|M_i(r)) = \exp\{c_i(r, M_i(r))\}$, from equation (5.2), for $i = 1, \ldots, S$; and $p(\bar{M}_g(r))$ here is proportional to a constant because we are assuming that a priori all subgroup network

structures are equally probable. The other term is defined as:

$$p(M_i(r)|\bar{M}_g(r)) \propto \exp\{-\lambda_{irg} d_{irg}\},$$

where d_{irg} is the SHD between $M_i(r)$ and $\bar{M}_g(r)$, *i.e.* the number of nodes that are the parents of node r only in one network: $M_i(r)$ or $\bar{M}_g(r)$. We reduce the number of hyper-parameters by assuming that the parents of node r are a priori equally likely to be shared between the subject i and subgroup g, *i.e.* $\lambda_{irg} = \lambda$ for all i, r, and g. The hyper-parameter λ is usually specified in a subjective manner. Oates (2013) suggested writing λ as a function of the probability of maintaining the status (present/absent) of the edge j between the individual network M_i and the subgroup network $\bar{M}_g$. That is,

$$p(j \notin M_i \Delta \bar{M}_g) = \frac{\exp\{-\lambda \times 0\}}{\exp\{-\lambda \times 0\} + \exp\{-\lambda \times 1\}} = \frac{1}{1 + \exp\{-\lambda\}}. \tag{6.2}$$

Here $A \Delta B$ denotes the set of elements contained in A but not in B plus elements contained in B but not in A (the proof of equation (6.2) can be seen in Appendix D.1). Therefore, the odds of an individual graph is the same as its subgroup graph regarding a particular edge is

$$O_\lambda := \frac{p(j \notin M_i \Delta \bar{M}_g)}{p(j \in M_i \Delta \bar{M}_g)} = \frac{\exp\{-\lambda \times 0\}}{\exp\{-\lambda \times 1\}} = \exp^{\lambda}.$$

For instance, considering $\lambda = 0.7$, the probability of maintaining edge status is almost twice the probability of not maintaining edge status between the subgroup and individual networks. This odds increases to about 20 and then 148 for $\lambda = 3$ and $\lambda = 5$, respectively.

Defining $p(M_i|\mathbf{y}) = \prod_{r=1}^{n} p(M_i(r)|\mathbf{y})$, the MAP estimator of individual networks is then

$$\hat{M}_i := \underset{M_i \in \mathbb{M}_i}{\arg\max}\, p(M_i|\mathbf{y}).$$

The individual network structure for subject i can thus be found using the scores in equation (6.1) and the MDM-IPA or the MDM-DGM.

MEMN: Subgroup Graphical Structures: $\bar{M}_1, \ldots, \bar{M}_G$

The subgroup network is found through the posterior probability of $\bar{M}_g(r)$, as follows:

$$\begin{aligned} p(\bar{M}_g(r)|\mathbf{y}) &= \sum_{\substack{M_i(r)\in\mathbb{M}_i(r)\\ i\in\mathbb{S}_g}} p(\bar{M}_g(r), M_{1^*}(r), \ldots, M_{S_g^*}(r)|\mathbf{y}) \\ &\propto \sum_{\substack{M_i(r)\in\mathbb{M}_i(r)\\ i\in\mathbb{S}_g}} \Big[p(\mathbf{y}|M_{1^*}(r), \ldots, M_{S_g^*}(r)) \times \\ &\quad p(M_{1^*}(r), \ldots, M_{S_g^*}(r)|\bar{M}_g(r)) \times p(\bar{M}_g(r)) \Big] \\ &\propto \prod_{i\in\mathbb{S}_g} \sum_{M_i(r)\in\mathbb{M}_i(r)} \Big[p(\mathbf{y}_i|M_i(r)) \times p(M_i(r)|\bar{M}_g(r)) \Big]. \end{aligned}$$

Again $p(\bar{M}_g(r))$ is considered as proportional to a constant. As $p(\bar{M}_g|\mathbf{y}) = \prod_{r=1}^{n} p(\bar{M}_g(r)|\mathbf{y})$, the subgroup network structure is found using these scores above and the MDM-IPA or the MDM-DGM, so that

$$\hat{\bar{M}}_g := \underset{\bar{M}_g\in\bar{\mathbb{M}}_g}{\arg\max}\, p(\bar{M}_g|\mathbf{y}).$$

MEMN: Group Graphical Structure: $\bar{M}$

The estimation of group network structure $\bar{M}$ using the individual networks $M_1, \ldots, M_S$ follows the same idea shown above for subgroup networks. Thus

$$p(\bar{M}(r)|\mathbf{y}) \propto \prod_{i=1}^{S} \sum_{M_i(r)\in\mathbb{M}_i(r)} \Big[p(\mathbf{y}_i|M_i(r)) \times p(M_i(r)|\bar{M}(r)) \Big], \tag{6.3}$$

and $p(\bar{M}|\mathbf{y}) = \prod_{r=1}^{n} p(\bar{M}(r)|\mathbf{y})$. The MAP estimator is

$$\hat{\bar{M}} := \underset{\bar{M}\in\bar{\mathbb{M}}}{\arg\max}\, p(\bar{M}|\mathbf{y}).$$

Comparing the MEMN with the IEMN in theory

In the IEMN, the individual networks are estimated independently, assuming that the subjects have different networks, and so the information of one subject is not used in the estimation of other subject. In contrast, the group network is estimated considering that all

subjects come from the same population and so they share the same graphical structure.

The higher λ, the more similar are the group network results, comparing the IEMN with the MEMN. This is due to the IEMN assumes that all subjects have the same graphical structure to estimate $\bar{M}$. Similarly, in the MEMN, the higher λ, the higher the score in which the distance between the individual and the group network is small, and so the more similar the individual graphs to each other.

The smaller λ, the more similar are the individual results $(M_1, \ldots, M_S)$, comparing the IEMN with the MEMN. When $\lambda = 0$, $p(M_i(r)|\bar{M}_g(r))$ is proportional to a constant, for $i = 1, \ldots, S$, and so $p(M_i(r)|\mathbf{y})$ is function of $c_i(r, M_i(r))$ (see equation (6.1)). Therefore, the estimated individual graphical structures using the IEMN are the same as the MEMN. In contrast, as λ increases, the scores are more penalised for individual networks that are more different from $\bar{M}_g$, increasing therefore the distance between the IEMN and the MEMN results.

Finally, when $\lambda = 0$, the scores of group network $\bar{M}$, using the MEMN, are proportional to a constant (see equation (6.3)), and so it is difficult to estimate the graphical structure. In this case, $p(\bar{M}(r))$ should be informative, and then the learning process is only based on the prior distribution of group network.

Note that the comments above about group network $\bar{M}$ also apply to the subgroup networks $\bar{M}_g$. Some properties cited above will be demonstrated in the next section.

6.3 The Application of Multiple Networks

We next use a rich fMRI study that has information from five different experimental conditions, hence called sessions: Session 1 is a resting-state condition; Session 2 is a motor condition in which individuals tapped their fingers; Session 3 is a visual condition in which individuals watched a movie; Session 4 and Session 5 are a combination between visual and motor condition, but the former is in a random way whilst in the latter, individuals tapped their fingers depending on random events in the movie.

Data were acquired on 15 subjects, and each acquisition consists of 230 time points, sampled every 1.3 seconds, with 2x2x2 mm^3 voxels. The FSL software[1] was used for preprocessing, including head motion correction, an automated artefact removal procedure (Salimi-

[1] `http://fsl.fmrib.ox.ac.uk`

Khorshidi *et al.*, 2014) and intersubject registration. We use 11 ROI's defined on 5 *motor* brain regions and 6 *visual* regions. The motor nodes used are Cerebellum, Putamen, Supplementary Motor Area (SMA), Precentral Gyrus and Postcentral Gyrus (nodes numbered from 1 to 5 respectively) whilst the visual nodes used are Visual Cortex V1, V2, V3, V4, V5 and task negative (v1+v2; nodes numbered from 6 to 11 respectively). The observed time series were computed as the average of BOLD fMRI data over the voxels of each of these defined brain areas.

6.3.1 Applying the Individual-Structure Approach

Firstly we applied the IS approach, modelling each subject using a search method, and then comparing the graphs of the brain connectivities across individuals. Using a weakly informative prior, with $n_0(r) = d_0(r) = 0.001$ and $\mathbf{C}_0^*(r) = 3\mathbf{I}_{p_r}$ for all r, the scores of all possible sets of parents for every node were found. The MDM-IPA was then used to discover the optimal graphical structure to explain the data from each subject. We assessed the intersubject consistency of the resulting networks by the prevalence of directed edges and by testing the null hypothesis of completely homogeneous connectivity over the network. Specifically, we estimated p_{ij}, the probability that an edge $i \to j$ exists, as the proportion $\hat{p}_{ij}$ of subjects with this particular edge between the identified regions. We used a one-sided Binomial test of $H_0 : p_{ij} = \pi$ versus $H_a : p_{ij} > \pi$, where π is the edge occurrence rate under homogeneity, sets equal to the average of $\hat{p}_{ij}$ over the 90 possible edges (see details in Appendix D.1).

Figure 6.3 shows $\hat{p}_{ij}$, but only for those edges with significant Binomial tests after false discovery rate correction (FDR; Benjaminin and Hocberg, 1995; Appendix D.1) at level $\alpha_{\text{FDR}} = 0.05$, where i indexes rows and j columns. The $\hat{p}_{ij}$ values for all edges can be seen in the Figure D6 in the Appendix D.3. The black horizontal and vertical lines divide the figure into four squares; the top left square represents the connectivity between motor brain regions, whilst the lower right square represents one between visual brain regions. Unsurprisingly, most of the connectivities are within these two squares. The two other squares represent *cross-modal* connections, between motor and visual regions, which are less prevalent.

The MDM was fitted for all subjects and sessions considering the graphical structure shown in Figure 6.3. The average of the discount factors over the subjects for each session is

given in Figure 6.2. The parameter δ for two motor areas, Precentral and Postcentral Gyrus (nodes 4 and 5), is above 0.9 for all sessions and so the system of these areas has a longer memory than others. However, other brain areas appear to be noticeably more volatile. In general it appears that visual nodes have a shorter memory than motor nodes. For instance, the average DF for nodes that have parents in Session 1 was 0.96 for motor nodes and was 0.80 for visual nodes. A possible reason is that the physical/sensory environment is much more constrained/static than the visual environment. For example, with this resting-state experiment, subjects were shown a screen with a fixation point. However, they were not explicitly asked to fixate. This might explain the greater perceptual variability in visual relative to sensory-motor areas.

Session 1, resting-state, has one of the smallest δ comparing with other sessions, except for Visual Cortex V1 (node 6) which has parents only in Session 1. One possible reason is that subjects do not do a specific activity in the resting-state experiment, being free to switch between different mental activities during the experiment. In contrast, Session 5, tapping depending on random events in the movie, has a long memory — they have one of the largest δ, except for Cerebellum, SMA and V4 areas (nodes 1, 3 and 9 respectively). An example of diagnostic analysis for this data is provided for a particular subject and resting-state session in Appendix D.2.

We analysed the results of the MDM-DGM algorithm across subjects as before. All values of $\hat{p}_{ij}$ can be seen in Figure D7 in Appendix D.3 whilst the significant connectivities are given in Figure 6.4. The nodes are ordered according to the expected flow of information in the brain, and thus, it is notable that we find significant edges between consecutive nodes. In general, Figure 6.3 also shows this pattern but less clearly for DAG constraints.

We also considered two other methods of estimating the functional connectivity: *full correlation* and *partial correlation* (Baba *et al.*, 2004; Marrelec *et al.*, 2006). For each node pair, per subject/session, we computed the full and partial correlation and converted each to a Z statistic with Fisher's transformation. For each node pair we tested the null hypothesis of mean zero (Fisher's transformed) correlation with a one-sample t-test (see details in Appendix D.1), corrected for multiplicity with FDR ($\alpha_{FDR} = 0.05$). Figure D8 and Figure D9, both in Appendix D.3, show the significant ($\alpha_{\text{FDR}} = 0.05$) full and partial correlation for every session, respectively. Note that these techniques provide symmetric results about the

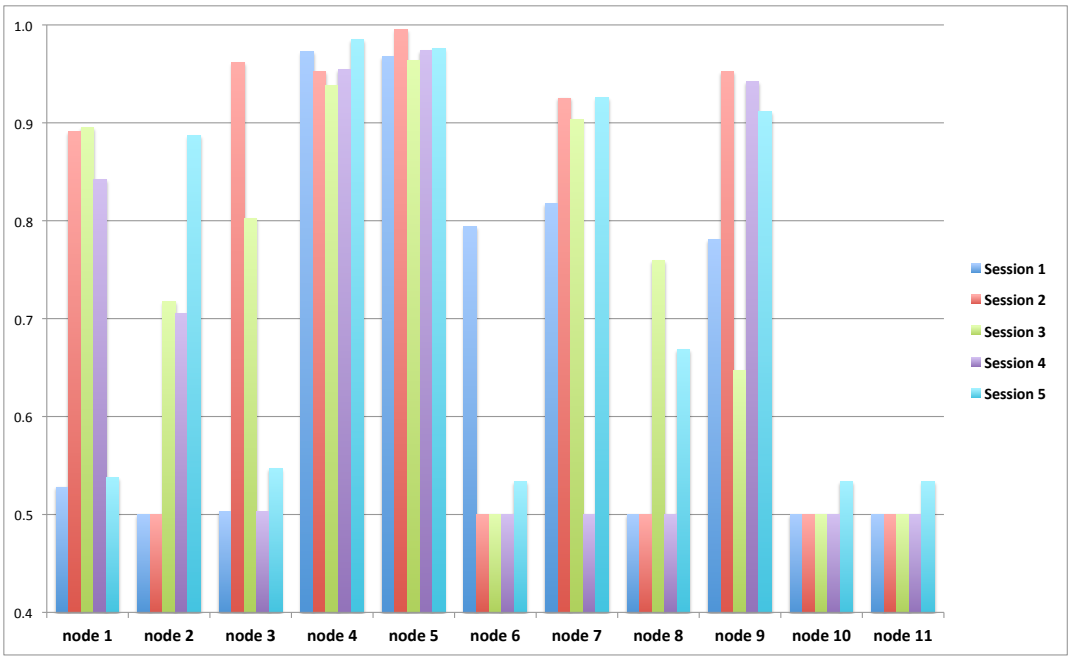


Figure 6.2: The average of the discount factors over the subjects using the graphical structure found in Figure 6.3 for each session. Nodes numbered from 1 to 5 are the motor regions: Cerebellum, Putamen, Supplementary Motor Area (SMA), Precentral Gyrus and Postcentral Gyrus, respectively. Nodes numbered from 6 to 11 are the visual regions: Visual Cortex V1, V2, V3, V4, V5 and task negative (v1+v2), respectively. Session 1 is a resting-state condition; Session 2 is a motor condition in which individuals tapped something; Session 3 is a visual condition in which individuals watched a movie; Session 4 and Session 5 are a combination between visual and motor condition, but the former is in a random way whilst in the latter, individuals tapping depending on random events in the movie.

principal diagonal. The vast majority of connections exist with high significance full correlation (Figure D8 in Appendix D.3). However, the connections with the strongest correlation (above 0.6) tended to be intra-modal as discussed above. As expected, the significant MDM edges are a subset of the significant partial correlations (Figure D9 in Appendix D.3). In short, while full and partial correlations do not account for nonstationarities nor represent a particular joint model, Figures 6.3 and 6.4 demonstrate that the application of the MDM gives scientifically plausible results and ones broadly compatible with other methods.

Another interesting analysis is to compare connections over different experimental conditions. In this way, the proportion of subjects who have a particular edge was compared for every pair of sessions. Using the McNemar test on paired proportions (see details in Appendix D.1) and p-values adjusted by FDR, there was no significant difference between the sessions for the MDM-IPA. However, for the MDM-DGM algorithm, Figure 6.5 shows the significant difference between the proportions of people who have a particular connection in a pair of sessions. When Session 1, resting-state, was compared with other sessions, most of the differences between connections was positive (with colours above the dark blue in the colour scale). So, in this case, connections existed in resting-state but not in other experimental conditions. In general most of the differences between the sessions occurred in the connections between visual nodes (in the lower right square). Figures D10 and D11 in Appendix D.3 give the significant differences of full and partial correlations for every pair of sessions, respectively. Overall the number of significant different connections was highest between Session 1, resting-state, and other sessions. Session 3, visual condition, was closest to Session 4, visual and motor conditions.

6.3.2 Comparing the MEMN with the IEMN in Practice

In this section we discuss the main differences between the IEMN and the MEMN, highlighted in the previous section, considering now a real application. In addition, we explore some new methods for determining λ and discuss the impact of its values on the results of multi-subjects analyses. We show that, depending on the chosen value of λ, the IEMN and the MEMN can provide completely different results. Recall that the space of λ parameter is $(0, \infty)$, *i.e.* it ranges from individuals believed to have different connectivity maps to hypotheses that they have the same one. The elicitation of λ, in theory, was discussed in Section 6.2.2.

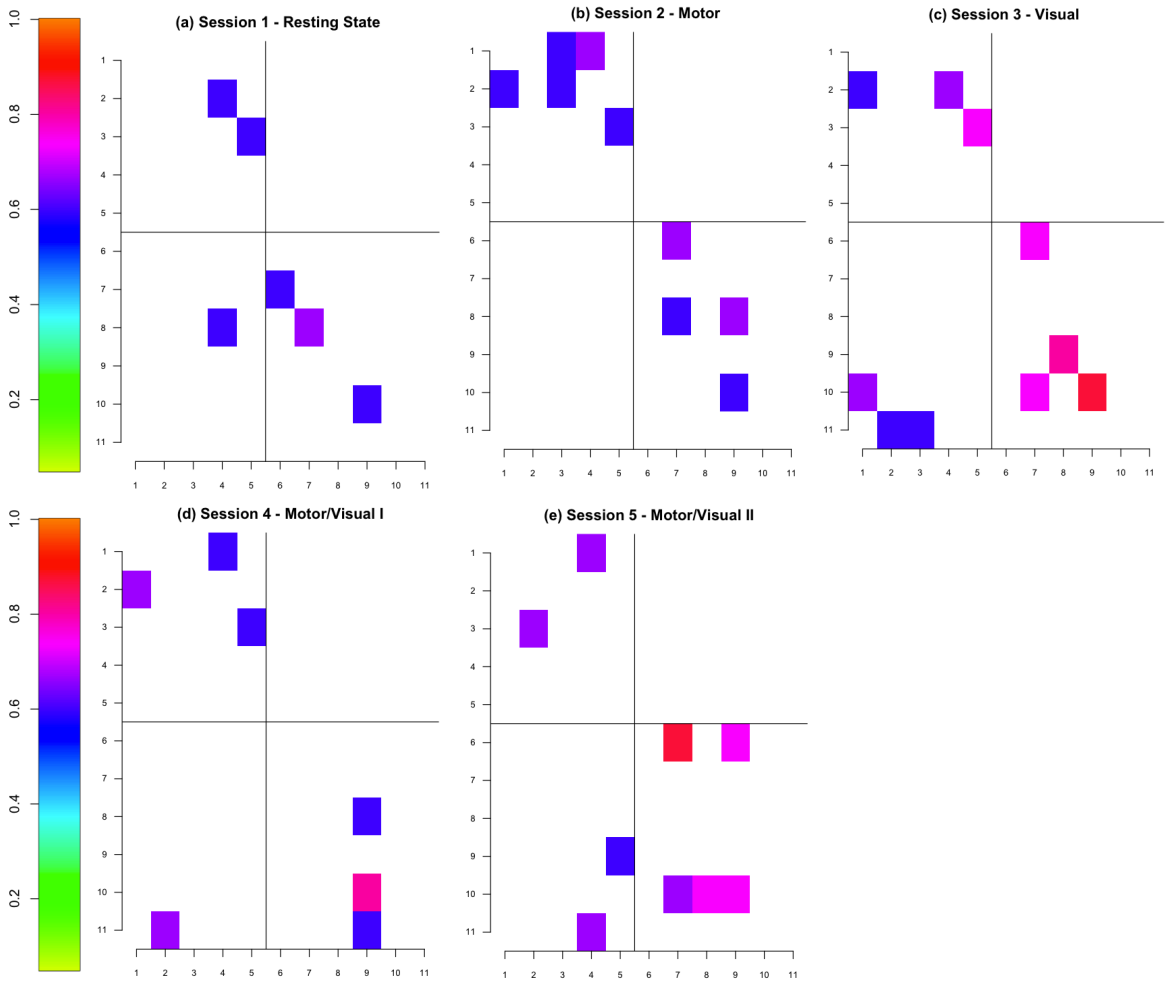


Figure 6.3: The proportion of subjects who have a particular edge $i \rightarrow j$, where i indexes rows and j columns, using the *MDM-IPA* per session, only for significant connectivities, $\alpha_{\text{FDR}} = 0.05$. Nodes numbered from 1 to 5 are motor regions, while nodes numbered from 6 to 11 are visual regions. The black horizontal and vertical lines divide the figure into four squares; the top left square represents the connectivity between motor brain regions, whilst the lower right square represents one between visual brain regions.Within each group, nodes are arranged according to the anticipated flow of information in the brain.

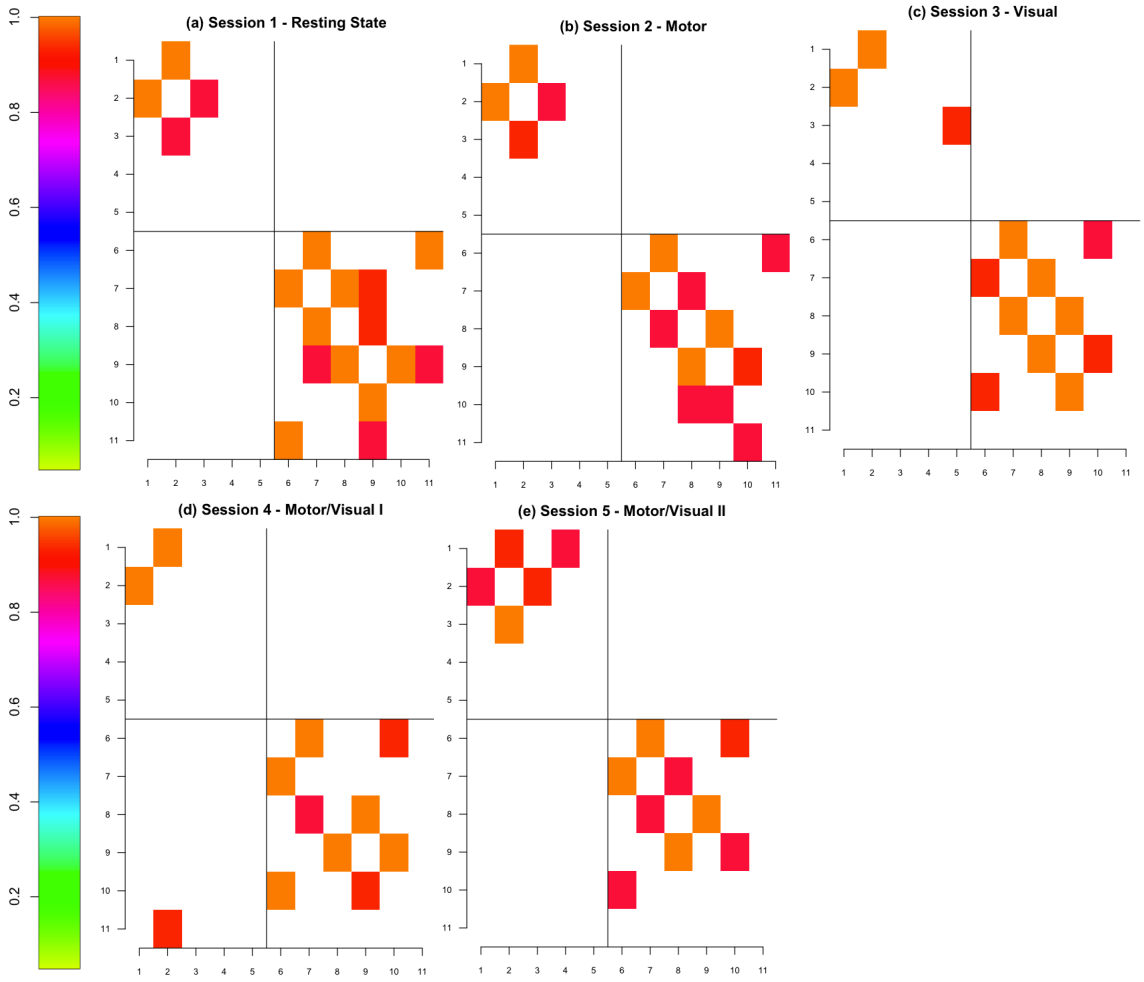


Figure 6.4: The proportion of subjects who have a particular edge $i \rightarrow j$, where i indexes rows and j columns, using the *MDM-DGM* per session, only for significant connectivities, $\alpha_{\text{FDR}} = 0.05$. Nodes numbered from 1 to 5 are motor regions, while nodes numbered from 6 to 11 are visual regions. The black horizontal and vertical lines divide the figure into four squares; the top left square represents the connectivity between motor brain regions, whilst the lower right square represents one between visual brain regions.Within each group, nodes are arranged according to the anticipated flow of information in the brain.

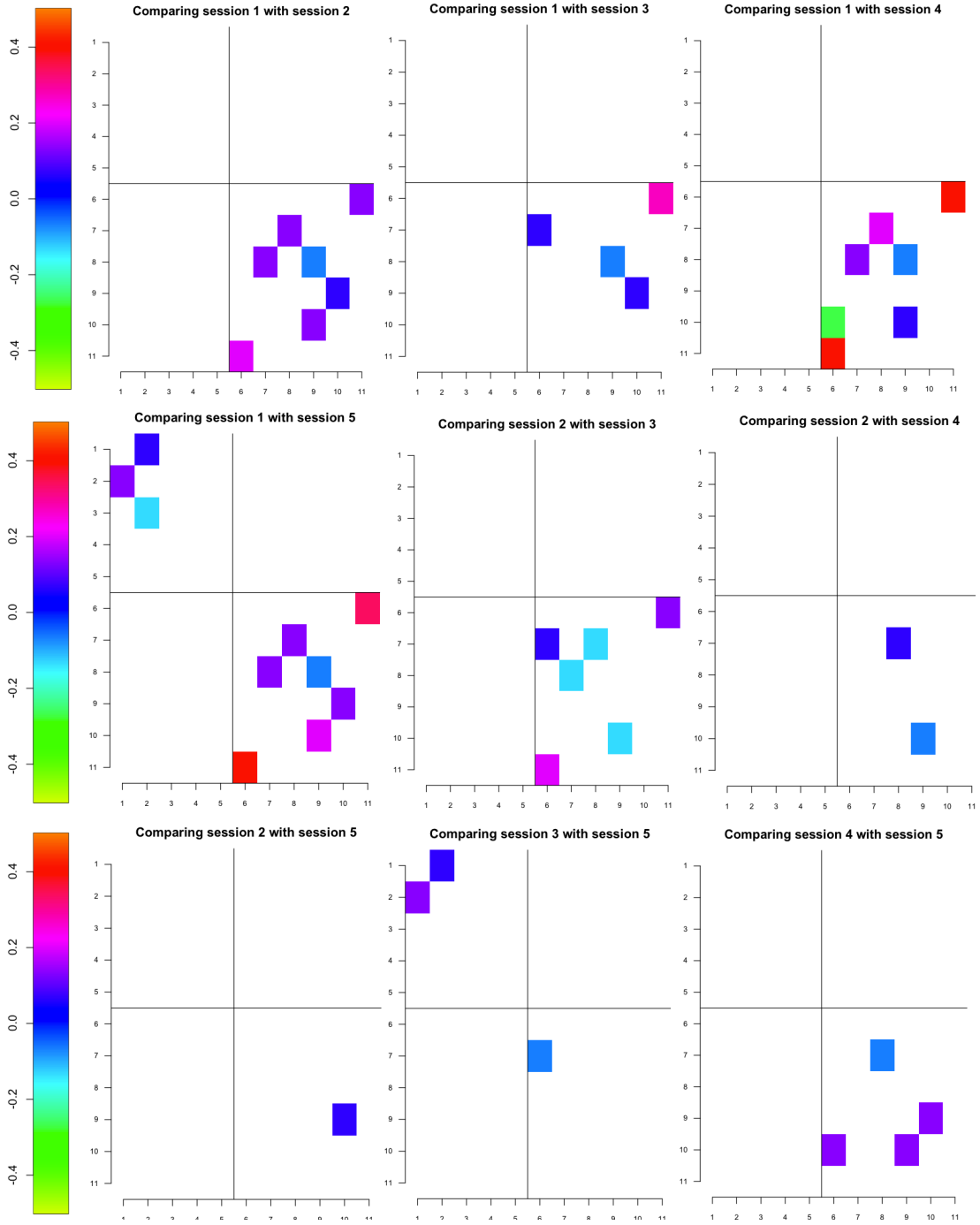


Figure 6.5: The significant difference of the proportion of subjects who have a particular connection $i \rightarrow j$ between two sessions using the *MDM-DGM algorithm*. Session 1 is a resting-state condition; Session 2 is a motor condition in which individuals tapped something; Session 3 is a visual condition in which individuals watched a movie; Session 4 and Session 5 are a combination between visual and motor condition, but the former is in a random way whilst in the latter, individuals tapping depending on random events in the movie.

Some possibilities of defining this parameter are (i) though a scientific belief statement (Oates, 2013), *e.g.*, as shown in Section 6.2.2, $\lambda = 0.7$ implies that the probability of maintaining edge status (absent/present) is almost twice the probability of not maintaining edge status between the group and the individual networks; (ii) maximising the LPL (or equivalently, maximising the scores $c_i(r, M_i(r))$ or $c(r, \bar{M}(r))$; (iii) maximising the posterior probability of individual networks, $p(M_i|\mathbf{y})$, or group networks, $p(\bar{M}|\mathbf{y})$; (iv) by cross-validation; or yet (v) considering the replications of the same subjects, when they are available. This fifth possibility to set λ will be discussed in Section 6.4.2, others we present below. We show here that different ways of estimating λ may provide different values of this parameter, and so divergent analysis results. Therefore, the chosen of the method used to estimate λ shall be in accordance with the objective of the study, as discussed below.

In this section, we also compare the MDM-IPA and the MDM-DGM. Here we show similar results to those found in the previous section, *e.g.* the graphs estimated by the MDM-DGM are usually denser than DAGs from the MDM-IPA, to accommodate for the possible cycles in the communication among brain regions. We also show that the methods described here can be used to compare data from different experimental conditions.

For the purposes described above, we are using an external validation study, in which the estimated individual networks were compared to predictive networks, *i.e.* networks estimated using the data from other $S - 1$ subjects. For simplicity, we are considering two levels: the individual and the group network, but, of course, this analysis can also be applied to subgroup networks as well.

Firstly the individual networks were estimated considering the IEMN and the MEMN, with $\lambda = 0.1, 0.7, 10, 100$ and 1000, using the MDM-IPA and the MDM-DGM. Then the predicted individual network for subject i was estimated considering the group analysis for all subjects, except subject i, using the same methods as before. The logBF and the SHD were used to compare methods. In addition, some analyses considered only the significant edges, *i.e.* let $\theta^*_{tij} = sm_{tij}/\sqrt{sC_{tij}}$, where sm_{tij} and sC_{tij} are the location and scale parameters of the smoothed distribution at time t, for edge j and subject i. A significant edge is then deemed to be one that has $\bar{\theta}^*_{ij} \geq 2$, where $\bar{\theta}^*_{ij}$ is the average of θ^*_{tij} over time.

Figure 6.6 shows the estimated and predicted graphical structures for subject 1 in the resting-state condition, considering all methods described above. As expected, the estimated

individual network found using the IEMN was similar to the one using the MEMN with small λ (see first and second column and first and third row). In contrast, as the predicted results were found using the methods of group network and as discussed above, the predicted network of IEMN was similar to one of the MEMN with large λ (see first and last column and second and last row).

These results are confirmed in Figure 6.7 that provided the average of the logBF comparing the IEMN with the MEMN ($\lambda = 0.1, 0.7, 10, 100, 1000$) over subjects and sessions, using the MDM-IPA (blue bars) and the MDM-DGM (orange bars), for estimated (left) and predicted (right) networks. Thus, the distance between the IEMN and the MEMN increased with the value of λ for individual networks and decreased for group networks. Note that we are here evaluating logBF based on the scores $c_i(r, M_i(r))$ which are the LPL for subject i when we are using the MDM-IPA, but they do not form the joint distribution of nodes for the MDM-DGM, as already discussed in Section 3.3.2. The interpretation of this measure is therefore rather fragile in the case of the MDM-DGM.

Unsurprisingly $\log \mathrm{BF} > 0$, because the IEMN network was found maximising the scores $c_i(r, M_i(r))$ for individual networks or $c(r, \bar{M}(r))$ for group network. The value of λ that maximised these scores was, therefore, 0.1 for individual networks and 1000 for group networks. This pattern shown in Figure 6.7 was also found in the analysis per session and considering only the significant edges.

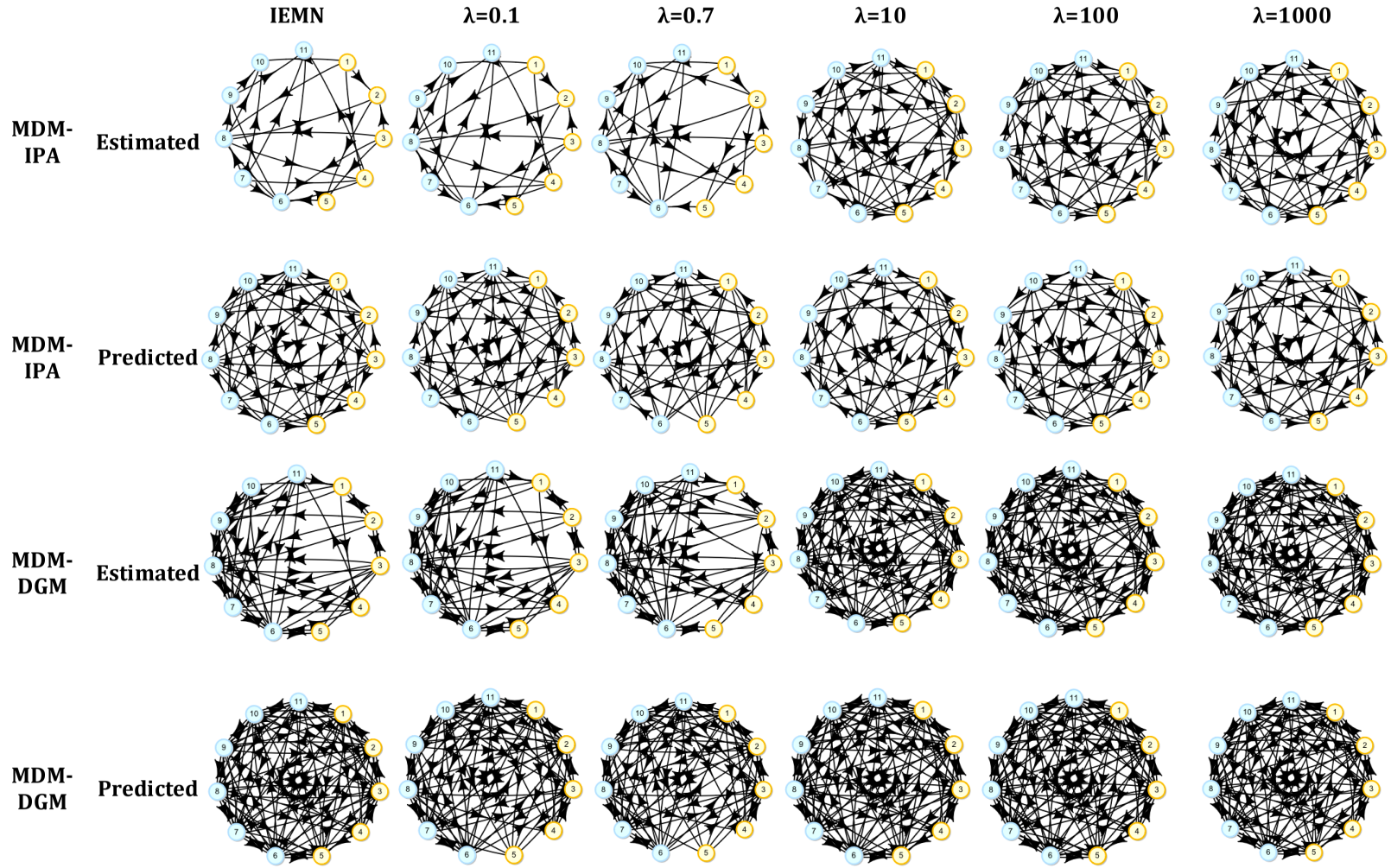


Figure 6.6: The estimated and predicted networks, using the MDM-IPA and the MDM-DGM, for subject 1 and resting-state condition (Session 1), considering the IEMN (the first column) and MEMN (from the second column) with $\lambda = 0.1, 0.7, 10, 100$ and 1000. The motor nodes used are Cerebellum, Putamen, Supplementary Motor Area (SMA), Precentral Gyrus and Postcentral Gyrus (yellow nodes numbered from 1 to 5 respectively) whilst the visual nodes used are Visual Cortex V1, V2, V3, V4, V5 and task negative (v1+v2; blue nodes numbered from 6 to 11 respectively).

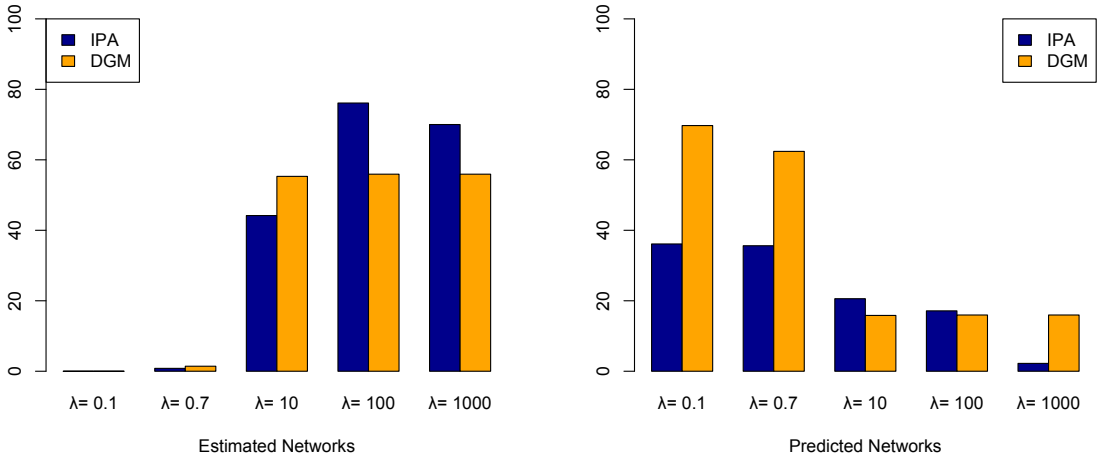


Figure 6.7: The average of logBF comparing the IEMN with the MEMN ($\lambda = 0.1, 0.7, 10, 100, 1000$) over subjects and sessions, using the MDM-IPA (blue bars) and the MDM-DGM (orange bars), for estimated (*left*) and predicted (*right*) networks.

The individual $c_i(M_i)$ and group $c(\bar{M})$ scores were evaluated as a function of the posterior probability of individual and group networks, by equation (6.1) and equation (6.3) respectively, *i.e.*

$$\begin{aligned} c_i(M_i) &:= \log p(M_i|\mathbf{y}) = \sum_{r=1}^{n} \log p(M_i(r)|\mathbf{y}), \\ c(\bar{M}) &:= \log p(\bar{M}|\mathbf{y}) = \sum_{r=1}^{n} \log p(\bar{M}(r)|\mathbf{y}). \end{aligned}$$

The value of λ that maximised these scores was 0.7 for both individual and group networks, and also for the MDM-IPA and the MDM-DGM (see Table 6.1).

λ	Individual		Group	
	IPA	DGM	IPA	DGM
0.1	20336.50	20851.97	19414.24	19426.27
0.7	20382.76	20904.73	19423.86	19507.02
10	19929.91	20731.38	18119.40	19348.62
100	18844.87	20729.59	14671.61	19347.24
1000	14349.80	20729.59	13499.95	19347.24

Table 6.1: The average of scores for individual networks, $c_i(M_i)$, and group networks, $c(\bar{M})$, over subjects and sessions, considering the learning network algorithms: the MDM-IPA and the MDM-DGM, and different values of λ.

Analysing the number of edges, the results also clearly depend on the individual and the group networks. We can see in Figure 6.6 that the higher λ, the denser were the estimated individual networks (first and third row) whilst the group networks got sparser

with the growth of λ (second and last row). Similar results are shown in Figure 6.8 (top) which shows the average difference between the number of edges of the IEMN and the MEMN, for individual (left) and group (right) networks. The MEMN provided denser graphs than the IEMN as the λ increased for individual networks whilst the graphs of the former was sparser than the latter, for group networks found by the MDM-DGM. Using the MDM-IPA and group networks, the IEMN provided denser graphs than the MEMN for $\lambda = 0.1$ and 0.7, but sparser for $\lambda \geq 10$, albeit this difference was small. In contrast, considering only significant edges (bottom), the IEMN provided mostly graphs slight denser than the MEMN. In general, the difference of the number of edges between the IEMN and the MEMN graphs is smaller for the MDM-IPA than for the MDM-DGM.

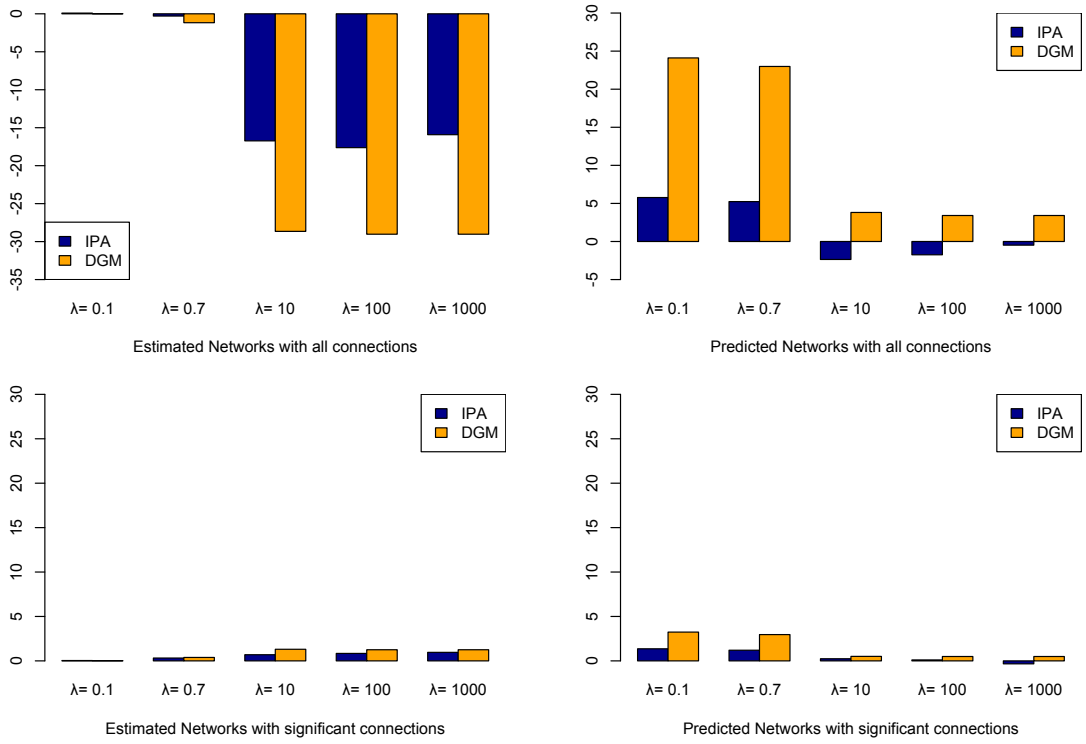


Figure 6.8: The average of the difference between the number of edges comparing the IEMN with the MEMN ($\lambda = 0.1, 0.7, 10, 100, 1000$) over subjects and sessions, using the MDM-IPA (blue bars) and the MDM-DGM (orange bars), for estimated (left) and predicted (right) networks, and considering all edges (top) and only significant edges (bottom).

Now we study which method provides graphical structures more similar over subjects. Figure 6.9 shows the average of the SHD between two estimated individual networks, over all pairwise subjects and all sessions, considering the IEMN and the MEMN ($\lambda = 0.1, 0.7, 10, 100, 1000$), using the MDM-IPA (blue bars) and the MDM-DGM (orange bars). Considering the complete individual graphical structure (Figure 6.9, left), the IEMN

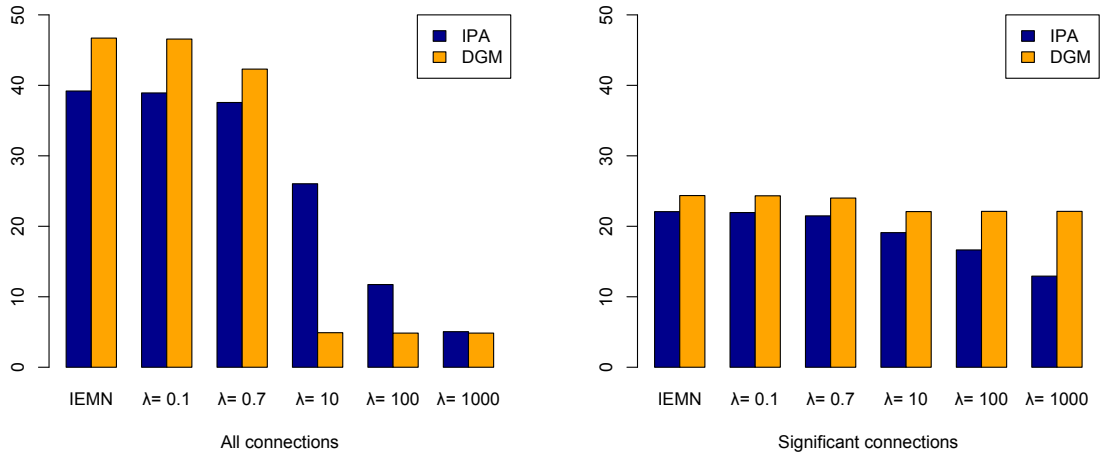


Figure 6.9: The average of the structural Hamming distance comparing two individual networks over all pairwise subjects and sessions, considering the IEMN and the MEMN ($\lambda = 0.1, 0.7, 10, 100, 1000$), using the MDM-IPA (blue bars) and the MDM-DGM (orange bars), for the all edges (left) and the only significant edges (right) of estimated networks.

and the small values of λ for the MEMN provided different results across subjects. This result is expected as long as the large λ implies to a similar structure between the individual networks (M_i's) and the group network ($\bar{M}$), and then, among the individual graphs. In this way, the MEMN with large λ is suitable for a homogeneous group. This conclusion is confirmed considering the estimated individual graphs, with only the significant connections (right figure), albeit the difference between the methods is faint.

The estimated and the predicted networks were compared using the logBF and the SHD, for the complete graph and considering only significant edges. Figure D12 in Appendix D.3 shows the average of these distances over subjects and sessions. We then computed the percentage of subjects in which the distance between the estimated and the predicted is the smallest when both networks belong to the same session. Figure 6.10 provides the average of this percentage of predicting correctly the network session, over all sessions, comparing the estimated to the predicted networks using the same method, *i.e.* the IEMN and the MEMN with $\lambda = 0.1, 0.7, 10, 100, 1000$, and comparing the estimated networks using the IEMN (*i.e.* individual graphs estimated independently) to the predicted networks using the MEMN for $\lambda = 0.1, 0.7, 10, 100, 1000$ (we called this l_λ). The chance of predicting the correct network session randomly is 1/5 sessions (dashed horizontal line in this figure). The green lines represent the 95% HPD intervals, considering a non-informative prior distribution, beta(1,1). In general, the IEMN and the MEMN with $\lambda = 1000$ provided one of the best results. For the MDM-IPA, the MEMN with $\lambda = 0.1$ had the highest percentage of predicting

the network session (around 40%) correctly. Overall the MDM-IPA predicted the network session more correctly than the MDM-DGM.

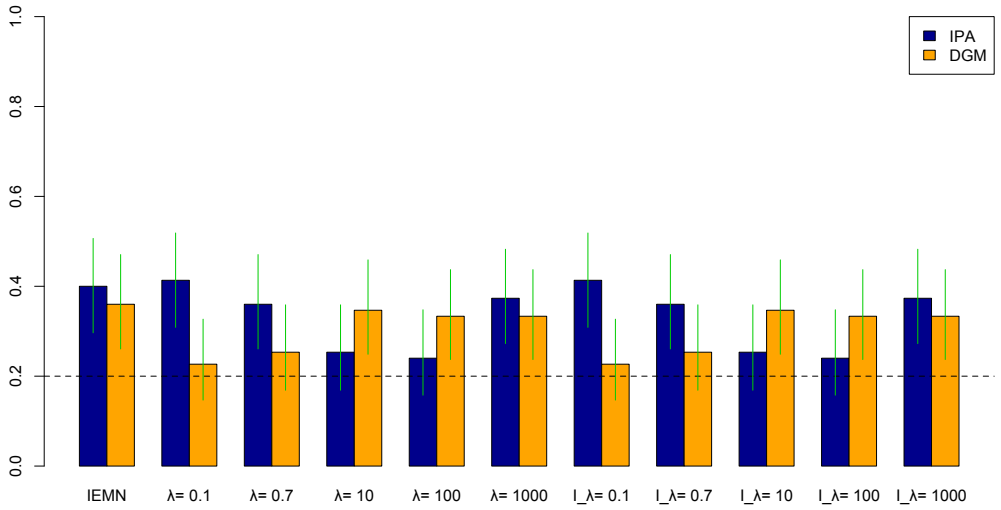


Figure 6.10: The average of the percentage of predicting correctly the network session, over all sessions, comparing the estimated to the predicted networks using the same method, *i.e.* the IEMN and the MEMN with $\lambda = 0.1, 0.7, 10, 100, 1000$, and comparing the estimated networks using the IEMN to the predicted networks using the MEMN for $\mathrm{l}\_\lambda = 0.1, 0.7, 10, 100, 1000$, using the MDM-IPA (blue bars) and the MDM-DGM (orange bars). The dashed horizontal line means the chance of predicting correctly the network session randomly. The green lines represent the 95% HPD intervals, considering a non-informative prior distribution, beta(1,1).

To compare sessions, we identified which sessions were better at predicting another particular session. For instance, Table 6.2 shows sessions (columns 3 and 5) that have the highest percentage of predicting the network of session cited in column 1, considering the method MEMN $\mathrm{l}\_\lambda = 1000$. Considering all methods and different values of λ, in general,

- Session 1, resting-state condition, better predicted the Session 2, motor condition;
- Session 3, visual condition, and Session 4, visual and motor (random) condition, better predicted each other;
- Session 3 also better predicted the Session 5, visual and motor condition;
- There was not a predominant session that predicted the Session 1, resting-state.

Note that these results are consistent with the conclusion of the previous section, where Session 1, resting-state, was responsible for the greatest difference among sessions, and Session 3, visual condition, is closer to Session 4, visual and motor conditions, than to other sessions.

Session	MDM-IPA		MDM-DGM	
	% right session	predictor session (%)	% right session	predictor session (%)
1-RS	40%	5 (33%)	47%	3 (27%) and 4 (27%)
2-Motor	47%	1 (33%)	33%	1 (27%)
3-Visual	33%	4 (40%)	40%	4 (33%)
4-Visual+Motor (random)	40%	3 (33%)	33%	1 (40%)
5-Visual+Motor	27%	3 (40%)	13%	4 (47%)

Table 6.2: The percentage of subjects who have a particular session chosen by the smallest value of the logBF comparing estimated network, using the IEMN, to the predicted network, using the MEMN with $\lambda = 1000$. Columns 2 and 4 show this percentage regarding the same session as in column 1, for the MDM-IPA and the MDM-DGM, respectively. Columns 3 and 5 give the session (and the correspondent percentage) that better predicts the session in column 1, whereas all other sessions, for the MDM-IPA and the MDM-DGM, respectively.

6.4 The Joint Estimation of Multiple Networks (JEMN)

In Chapter 5 we described a GS approach, which consisted of grouping homogeneous individuals according to their connectivity patterns, and then using the CS approach to find the subgroup networks. However, Ramsey *et al.* (2010) talked about the triangulation problem in their CS approach, *i.e.* the appearance of a direct link between two variables that actually have only an indirect causal relation. Therefore, Ramsey *et al.* (2010) handled this situation including a penalty function in their search algorithm. The triangulation problem also seems to exist here with the CS approach. Using both synthetic and real datasets, the result of this approach was that all nodes were connected with each other. Moreover, Section 4.2 showed that the number of false positive (FP) connections increased with the number of nodes, albeit the FP connectivity strengths were close to zero most of the time. In group analysis, this problem seems to be worse because group networks tend to be dense to be able to model the differences between subjects. For instance, Figure 6.6 shows that the estimated graphs (individual networks) are sparser than the predicted graphs (found by group analysis).

Search network processes that penalises dense networks have been studied in literature, *e.g.* L_1 penalties in graphical LASSO (Mohan *et al.*, 2012; Danaher *et al.*, 2014), or approaches based on non-local priors (Consonni and La Rocca, 2010). Based on the MEMN, Oates *et al.* (2014) developed the Joint Estimation of Multiple Networks (JEMN), using a penalty function in order to estimate sparse DAGs. The JEMN also obtains a MAP estimate for all individual and group DAGs simultaneously, and allows us to estimate relationships

between the subjects themselves. Here we provide for the first time an application of this method to a real experimental data.

6.4.1 A Statistical Model for Joint Multi-Subject Analysis

Exact estimation of multiple DAGs

The score function that the JEMN uses to search individual networks is evaluated based on the posterior conditional distribution of $(M_1, \ldots, M_S)$ given an undirected network A, *i.e.*

$$p(M_1, \ldots, M_S|\mathbf{y}, A) \propto \left(\prod_{i=1}^{S} p(\mathbf{y}_i|M_i) \right) \times p(M_1, \ldots, M_S|A). \tag{6.4}$$

The first term is $\exp\left(\sum_{i=1}^{S}\sum_{r=1}^{n} c_i(r, M_i(r))\right)$, given by equation (5.2). The observed variables are independent of the matrix A given $(M_1, \ldots, M_S)$. The matrix A is defined as the adjacency matrix of a network whose nodes consist of subjects whilst edges represent the similarity between them. When A is complete, the JEMN assumes exchangeability so that any DAG M_i is equally likely a priori to be similar to any other DAG M_j $(i \neq j)$, and so all subjects form a homogenous group. Such an exchangeability assumption is implicit in much of the recent literature on multiple graphical models (Mohan *et al.*, 2012; Penfold *et al.*, 2012; Danaher *et al.*, 2014). However, exchangeability will be inappropriate when the collection of subjects contains nontrivial structure, such as subgroups, that correspond to differential neural connectivities, as discussed above. For instance, as a result of a cluster analysis, A can be defined as the edge $i - j$ exists if only if the subjects i and j belong to the same subgroup. See a representation of the JEMN in Figure 6.11.

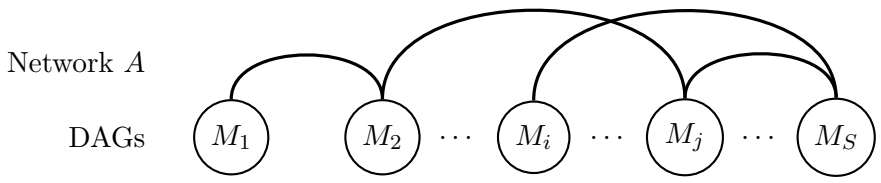


Figure 6.11: Individual networks: $M_1, \ldots, M_S$, with relationships between graphs encoded by an undirected network A.

The second term of equation (6.4) is the *multiple DAG prior* and is written as

$$p(M_1, \ldots, M_S|A) \propto \left(\prod_{(i,j)\in A} q(M_i, M_j) \right) \times \left(\prod_{i=1}^{S} m(M_i) \right),$$

where q is a similarity function defined as

$$q(M_i, M_j) \propto \prod_{r=1}^{n} \exp\{-\lambda d_{irj}\},$$

and d_{irk} is specified as before *i.e.* the SHD between $M_i(r)$ and $M_j(r)$. The function $m(M_i)$ provides an adjustment for the fact that the size of the DAG space grows super-exponentially with the number n of nodes. The JEMN follows Scott and Berger (2010) and controls multiplicity using the default correction

$$\log m(M_i) \quad = \quad \sum_{r=1}^{n} m_r(M_i(r)),$$

$$m_r(M_i(r)) = \begin{cases} -\log \binom{n}{|M_i(r)|} & \text{if } |M_i(r)| \leq d_{\max}, \\ -\infty & \text{otherwise;} \end{cases}$$

where $|M_i(r)|$ is the indegree of node r in M_i, and $d_{\max}$ is a fixed upper bound on the indegree of nodes in DAG that encodes prior knowledge on the support of the graphical models (*e.g.* Hill *et al.*, 2012).

Now the MAP estimator of all individual networks can be found simultaneously as

$$(\hat{M}_1, \ldots, \hat{M}_S)|A := \underset{M_1,\ldots,M_S \in \mathbb{M}_i^S}{\arg\max}\, p(M_1, \ldots, M_S|\mathbf{y}, A),$$

Or equivalently as

$$(\hat{M}_1, \ldots, \hat{M}_S)|A := \underset{M_1,\ldots,M_S \in \mathbb{M}_i^S}{\arg\max} \sum_{r=1}^{n} \left[\sum_{i=1}^{S} (c_i(r, M_i(r)) + m_r(M_i(r))) - \lambda \sum_{(i,j)\in A} d_{irj} \right], \quad (6.5)$$

where the space of all possible joint individual networks is $\mathbb{M}_i^S = \mathbb{M}_1 \times \ldots \times \mathbb{M}_S$.

Unknown similarity matrix A

When the network A is unknown, a hyperprior distribution is given by

$$p(A) \propto \prod_{(i,j)\in A} \exp(\eta_{ij}).$$

This hyperprior distribution has the effect of deterring sparsity in the network A, leading to increasing the regularisation between DAGs and a more conservative estimate of between-subject variability. Thus, the constants η_{ij} can be used to encode which subjects are more likely to share similar connectivity, based on ancillary covariates such as age, gender, disease status, etc. For example one could exploit $\eta_{ij} = \eta^{|\text{age}(i)-\text{age}(j)|}$, for some $\eta > 0$, that encourages sharing of graph structure among j and i subjects of similar ages. However, it is often practical to assume that all pairs of subjects are a priori equally likely to share similar graph structure, *i.e.* $\eta_{ij} = \eta$ for all i, j. Prior elicitation in this reduced class of models, therefore, requires the specification of hyperparameters λ and η.

The MAP estimator of all individuals networks and also the network A in the space $\mathcal{A}$ can be written as

$$\begin{aligned}(\hat{M}_1, \ldots, \hat{M}_S, \hat{A}) &:= \underset{\substack{M_1,\ldots,M_S\in\mathbb{M}_i^S \\ A\in\mathcal{A}}}{\arg\max}\ p(M_1, \ldots, M_S, A|\mathbf{y}) \quad \text{or} \\ &:= \underset{\substack{M_1,\ldots,M_S\in\mathbb{M}_i^S \\ A\in\mathcal{A}}}{\arg\max}\ p(M_1, \ldots, M_S|\mathbf{y}, A) \times p(A). \qquad (6.6)\end{aligned}$$

Note that when all elements of the matrix A are zero, equation (6.6) is equal to equation (6.5) with the last term being zero. This means that $(M_1, \ldots, M_S)$ are estimated independently as in the IEMN. Moreover, when $\eta = 0$, the prior distribution of A is proportional to a constant. Then the score function in equation (6.6) becomes the same as equation (6.5). Moreover, the hyperparameter η controls the density of A, or the homogeneity of the group of subjects. For instance, suppose three subjects and, by convention, the adjacent matrix A is upper triangular, so that $A = \begin{pmatrix} 1 & 1 & 0 \\ - & 1 & 0 \\ - & - & 1 \end{pmatrix}$. Therefore, the first two subjects form a homogeneous subgroup. Considering now that the entire group is homogeneous, *i.e.* $A^* = \begin{pmatrix} 1 & 1 & 1 \\ - & 1 & 1 \\ - & - & 1 \end{pmatrix}$, then $\log p(A^*) - \log p(A) \propto 3\eta - \eta = 2\eta$, and so the higher the value of η, the higher the chance that A has more edges. We also show this result using real data in Section 6.4.2.

Estimating the subgroup networks

Below we extend the JEMN to accommodate the estimation of group or subgroup networks. The dimension of matrix A is now $(S+G)\times(S+G)$, where $\bar{M}_1,\ldots,\bar{M}_G$ corresponding to nodes $S+1,\ldots,S+G$. Here the matrix A has two assumptions: (i) there are no edges between nodes i and j, whenever $i<j\leq S$ or $S+1\leq i<j$, and (ii) there is only one edge incident at each of node i for $1\leq i\leq S$. Figure 6.12 gives a representation of the JEMN considering G subgroups. For instance, after applying the cluster analysis as shown in Chapter 5, the matrix A can be defined as the edge $i-j$ exists if only if subject i belongs to subgroup $j-S$, for $1\leq i\leq S$ and $S+1\leq j\leq S+G$.

Generalising the distribution of networks given in equation (6.4), we consider the score function as

$$\begin{aligned} p(M_1,\ldots,M_S,\bar{M}_1,\ldots,\bar{M}_G|\mathbf{y},A) &\propto \left(\prod_{i=1}^{S} p(\mathbf{y}_i|M_i)\right) \\ &\times \quad p(M_1,\ldots,M_S,\bar{M}_1,\ldots,\bar{M}_G|A), \end{aligned}$$

where

$$\begin{aligned} p(M_1,\ldots,M_S,\bar{M}_1,\ldots,\bar{M}_G|A) &\propto \left(\prod_{(i,j)\in A} q(M_i,M_j)\right)\times\left(\prod_{i=1}^{S} m(M_i)\right) \\ &\propto \left(\prod_{(i,j)\in A}\prod_{r=1}^{n}\exp(-\lambda d_{irj})\right)\times\left(\prod_{i=1}^{S} m(M_i)\right). \end{aligned}$$

Therefore, the individual and subgroup networks are estimated simultaneously through

$$\{\hat{M}_1,\ldots,\hat{M}_S,\hat{\bar{M}}_1,\ldots,\hat{\bar{M}}_G\}|A := \underset{\substack{M_1,\ldots,M_S\in\mathbb{M}_i^S\\ \bar{M}_1,\ldots,\bar{M}_G\in\mathbb{M}_g^G}}{\arg\max}\; p(M_1,\ldots,M_S,\bar{M}_1,\ldots,\bar{M}_G|\mathbf{y},A),$$

where the space of all possible joint subgroup networks is $\bar{\mathbb{M}}_g^G=\bar{\mathbb{M}}_1\times\ldots\times\bar{\mathbb{M}}_G$.

The JEMN also allows a cluster analysis procedure, considering A as unknown, but given the number of subgroups G. In this case, the similarity between subjects is assessed indirectly by the SHD rather than the log Bayes factor separation provided in Chapter 5. Note that, in practice, the same procedure as before can be used to search networks, *i.e.*

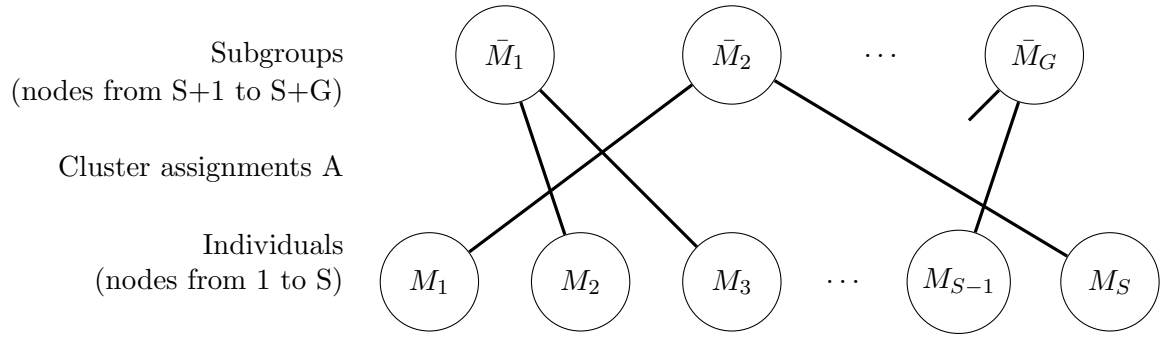


Figure 6.12: An example of clustering using the JEMN: $(\bar{M}_1, \ldots, \bar{M}_G)$ are data-generating graphs which correspond to nodes from $S+1$ to $S+G$ of network A, whilst $(M_1, \ldots, M_S)$ are individual exemplars and correspond to nodes from 1 to S.

equation (6.5) for known A and equation (6.6) for unknown A, limiting the space of A by assumptions (i) and (ii) provided above.

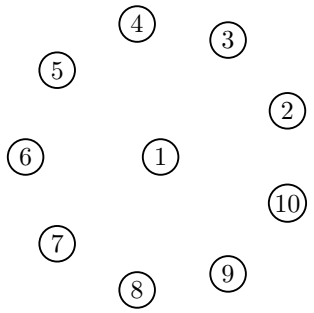
6.4.2 The Application of the JEMN into a Real FMRI Data

Our real fMRI datasets consist of a resting-state experiment with four replications obtained, under identical laboratory conditions, for each of six unrelated subjects from the Human Connectome Project (Van Essen *et al.*, 2013). Scans were acquired on each subject while they were in a state of quiet repose; data from one 15 minute session were used, with a spatial resolution of $2\times2\times2$ mm^3 and a temporal resolution of 0.7 secs, see Smith *et al.* (2013) for full details. After correcting for head motion, all data were registered to a common reference atlas space and 100-dimensional ICA was conducted on the temporally concatenated data. The result of this ICA was 100 spatial modes (common to all subjects) and 100 corresponding temporal modes (subject-specific); at this high dimension, the 100 spatial modes are sparse and spatially compact (though possibly bilaterally symmetric) and so essentially provide a data-driven parcellation of the brain. Hierarchical clustering was then used on the time series data and the 10-mode cluster corresponding to motor cortex was selected for study here. Thus, our data consists of 10 nodes and 1200 time points for each subject. Figure 6.13 shows the approximate description of each node. Note that node 4 was spatially diffuse and difficult to characterise, and thus is likely to be an artifactual component.

For all examples in this section we made the subjective choice $d_{\max} = 3$ that reflects the degree of connectivity observed in previous literature (*e.g.* Ramsey *et al.* 2010). Independent estimation for the subject-specific DAGs M_i, based on the IEMN, the MDM-IPA

Node Number	Symmetry	Summary
1	Bilateral	Motor:hand/face
2	Bilateral	Sensory:All-but-face
3	Bilateral	Motor:All-but-face
4	Bilateral	UNKNOWN
5	Left Dominant	Sensorimotor: L Hand+Arms
6	Right Dominant	Sensorimotor: R Hand+Arms
7	Bilateral	Sensory: Trunk-to-feet
8	Bilateral	Sensory: Face
9	Bilateral	Auditory
10	Bilateral	Sensorimotor:All-but-face - Sensory:Face

(a) Neural regions



(b) Vertex layout

Figure 6.13: Illustrative resting-state fMRI dataset. We consider ten spatial nodes as described in (a), each having a corresponding time series for each subject. All graphs that we present will adopt the vertex layout shown in (b).

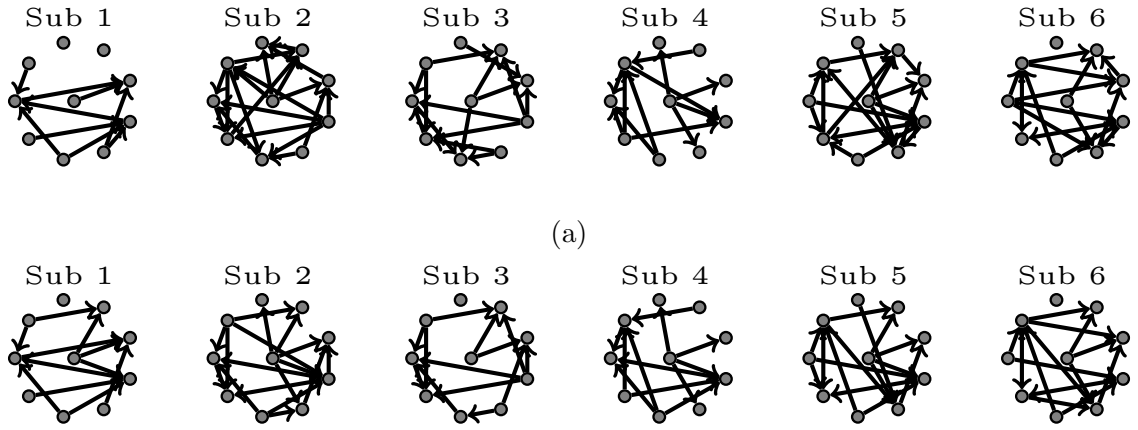


Figure 6.14: Networks for six subjects estimated using the MDM-IPA applied to one replication and for (a) each subject separately, (b) for all subjects jointly with regularity hyperparameter $\lambda = 4$ and considering A complete. The graphs in (b) are 23% more similar compared to the graphs in (a), as explained in the main text. Figure 6.13 (b) provides a key.

and only one replication, yields graphs that display high between-subject variability (Figure 6.14(a)). This is unexpected on scientific grounds, and likely reflects the lack-of-robustness and small sample bias that are often associated with graphical analyses.

Estimating the hyperparameter λ using replications

In order to establish how much regularisation is required for our illustrative fMRI dataset, we performed retrospective inspection of the posterior. Specifically, we performed exact estimation of the JEMN based on four technical replicate datasets obtained from the first two subjects. To elicit a suitable value for the regularity parameter λ, we fixed the population structure A such that $(k, l) \in A$ if and only if datasets k and l were both

technical replicates derived from the same subject (Figure 6.15). This corresponds to placing an exchangeability assumption on the technical replicates, but prohibiting the sharing of information between subjects. We then computed the total SHD between all pairs of DAGs that are technical replicates (Figure 6.16).

Recent studies (*e.g.* Ringach, 2009) indicate that the notion of resting-state is poorly defined and can correspond to several contrasting neurological activity profiles; we would therefore not expect to obtain identical DAGs under a replication experiment that is unable to control for the precise nature of the resting-state (*i.e.* we should have $\lambda < 17$, the point at which all DAGs become identical in Figure 6.15). In terms of the total SHD between replicates, as might be expected, Figure 6.16(a) shows that this distance decreased as λ increased (as shown above for the MEMN). Below for illustration we focus on one such value, $\lambda = 4$, that attributes approximately 50% of variability between technical replicates to extrinsic noise resulting from the experimental design. Examination of the Bayes factor as a function of λ provides a second diagnostic to assist with elicitation. In this case the value $\lambda = 4$ scores considerably better compared to the alternative that assigns the same DAG to all replicate datasets (logBF ≈ 900, Figure 6.16(b)).

Learning individual DAGs

Based on the elicitation $\lambda = 4$ and one replication per subject, firstly we employed the JEMN under the exchangeability assumption that A is the complete network (equation (6.5)). Results in Figure 6.14(b) demonstrate that the estimated DAG structure is substantially more regular than our original estimate obtained using independent inference (Figure 6.14(a)), with a 23% decrease in total SHD between DAGs, and can be expected more closely to represent the true subject-specific neural connectivity patterns. We note however that the validation of inferred connectivity remains extremely challenging (*e.g.* Stein *et al.*, 2007).

Note that here we simply consider one dataset per subject, for limiting scope, but the methodology presented above naturally accommodates data aggregation. For instance, the dimension of matrix A can be defined as $(4\times S+S+1)\times(4\times S+S+1)$, and the edge (i, j) exists (i) for $i < j \leq 4 \times S$ and datasets i and j are replications of the same subject (edges between *replications* and *individuals* in Figure 6.17); (ii) for $4\times S+1 \leq i \leq 4\times S+S$ and $j = 4\times S+S+1$ (edges between *individuals* and *group* in Figure 6.17). Nodes from 1 to $4 \times S + S + 1$ of A

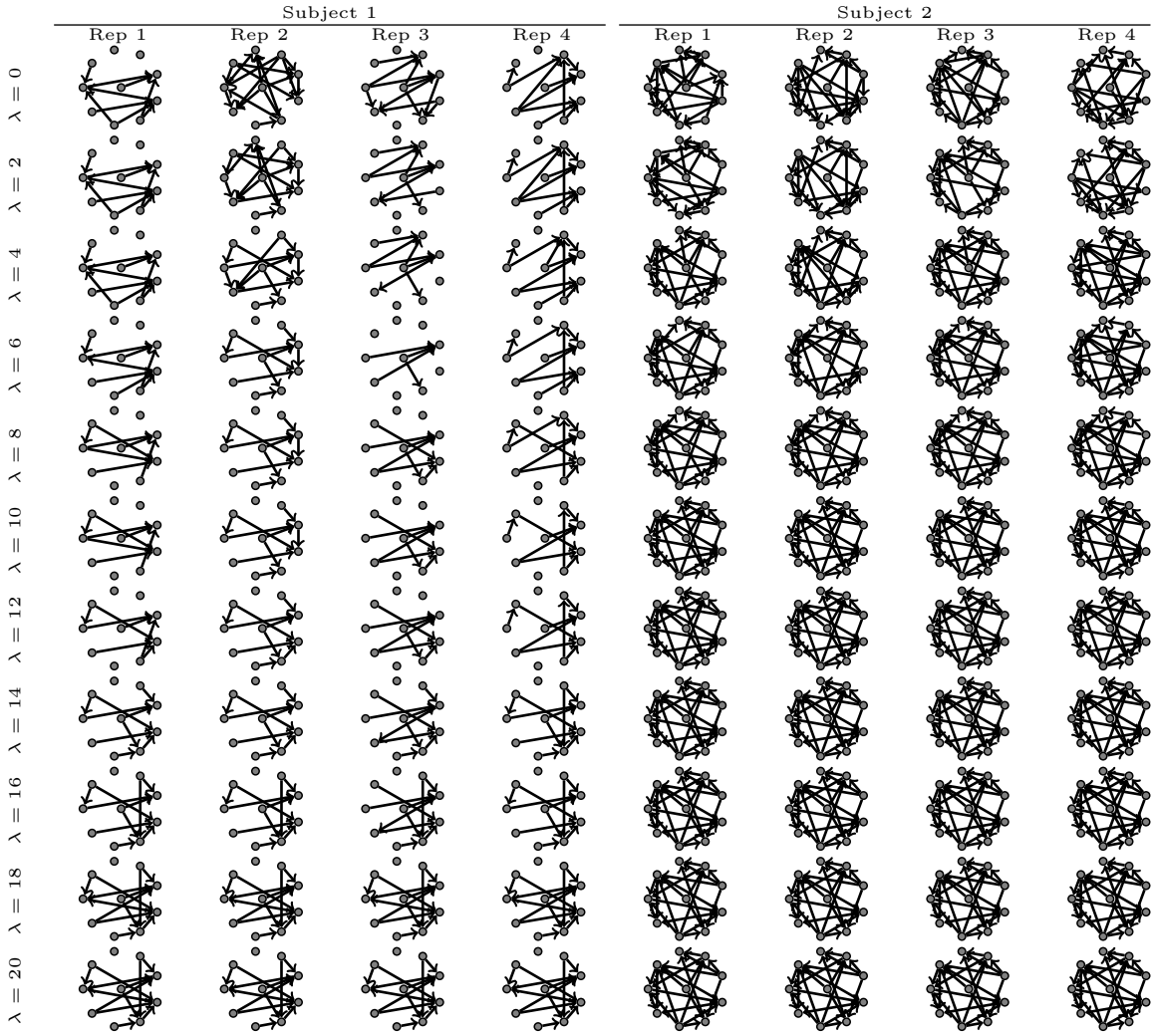
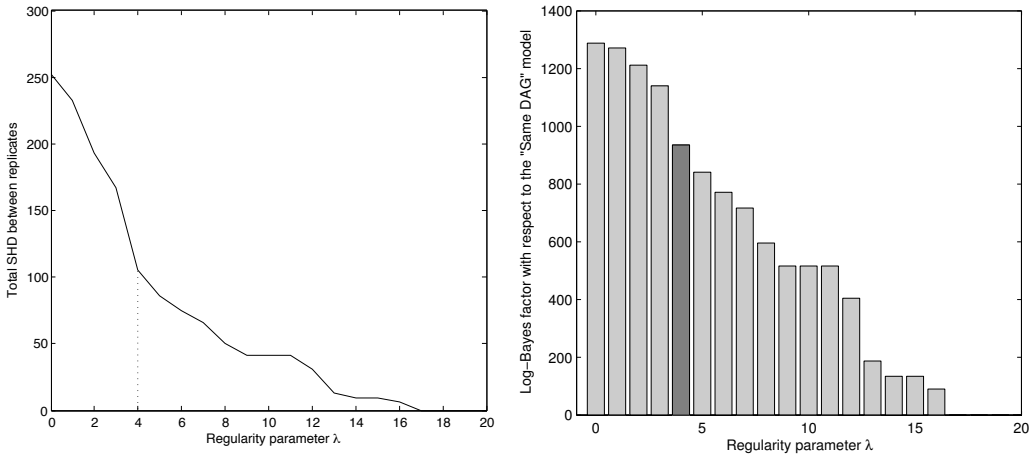


Figure 6.15: Eliciting a value for the regularity parameter λ based on technical replicate data and retrospective inspection of the posterior. Here two subjects each provided four technical replicate datasets. The DAGs shown are the JEMN estimates for varying λ, such that replicates were assumed to be exchangeable, but subjects were treated independently. As λ is increased the DAGs corresponding to technical replicates become more similar.

correspond respectively to $(M_{11}, \ldots, M_{41}, \ldots, M_{1S}, \ldots, M_{4S}, M_1, \ldots, M_S, \bar{M})$, where M_{ij} is the graph structure for the replication i of subject j, and M_i and $\bar{M}$ are defined as before.

As discussed above, the scientific motivation for multi-subject analysis is typically to elucidate differential connectivity between subjects, either in a purely unsupervised context for exploratory investigation, or in a supervised context to determine whether certain features of connectivity are associated with auxiliary covariates of interest such as disease status. In these cases a statistical model that assumes exchangeability between subjects may be inappropriate and "regularise away" the differential connectivity that is of interest. We therefore proceed to jointly estimate both individual networks and the network A that



(a) (b)

Figure 6.16: Eliciting a value for the regularity parameter λ. (a) Here we plot the total SHD between DAGs corresponding to technical replicates against the regularity parameter λ. The dashed line indicates the value $\lambda = 4$, that attributes approximately 50% of variability between replicates to extrinsic noise resulting from the experimental design. (b) Comparing the Bayes factor corresponding to model with a particular value of λ against the model that assumes all replications have the same DAG.

describes relationships between the subjects themselves (equation (6.6)).

Elicitation of the hyperparameter η (that controls density of the network A) was again performed by retrospective inspection of the posterior, requiring (i) a moderate amount of similarities between subjects, motivated by expectation that connectivity should not differ substantially between subjects, and (ii) a moderate amount of heterogeneity between subjects, since we aim to highlight any potential differences between the neural connectivity of different subjects. Results in Figure (6.18) demonstrate that for $\eta = 60$ the six subjects are regularised into three distinct subgroups $\{1, 4\}$, $\{2, 3\}$, $\{5, 6\}$, whilst for the higher value $\eta = 70$ the subjects are regularised into two distinct subgroups $\{1, 2, 3, 4\}$, $\{5, 6\}$. (When $\eta = 80$ the network A is complete and subject-specific DAGs coincide with Figure 6.14 (b)). The subgroups are formed according to hyperparameter η (see equation (6.6)) and minimising the SHD between subjects belonging to the same subgroup (see equation (6.5)). Therefore, the number of connectivities that coincide between DAGs is expected to be higher for subjects belonging to the same subgroup than for subjects from different subgroups.

Examination of the Bayes factor as a function of η demonstrates that the values $\eta = 60, 70$ are considerably better compared to the DAGs obtained under an exchangeability assumption (logBF $\approx$ 200, 180 respectively). This suggests that hierarchical group and subgroup structure may be present among the subjects at the level of neural connectivity

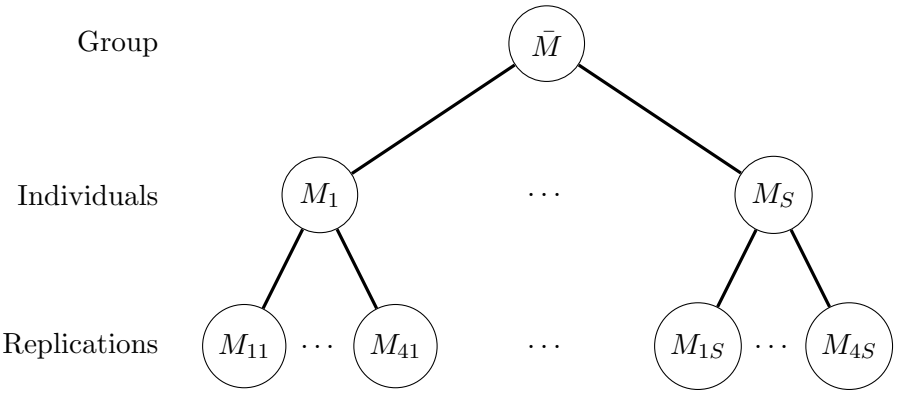


Figure 6.17: An example of the JEMN considering 4 replications for each subject. M_{ij} is the graph structure for the replication i of subject j; M_i is the graph structure for subject i; and $\bar{M}$ is the group network.

and provides evidence against exchangeability of the subjects.

Learning subgroup DAGs

Finally, we illustrate an alternative and novel approach to group individuals according to their similarities, and learning the individuals and subgroups DAGs simultaneously between subjects. Here we applied the JEMN to the six subjects using $G = 2$ clusters (Figure 6.19 (a)) and $G = 3$ clusters (Figure 6.19 (b)). The optimal cluster assignment with $G = 3$ recovers the three distinct subgroups $\{1, 4\}$, $\{2, 3\}$, $\{5, 6\}$ and with $G = 2$ was $\{1, 2, 3, 4\}$, $\{5, 6\}$ that were obtained above via joint estimation of A. This analysis reinforces, via an alternative route, the conclusion that models for the subjects in this dataset should not assume the exchangeability of subjects. Contrary to several methods that assumes exchangeability, as discussed mainly in the Section 5.1, *e.g.* the VST and the CS approaches (Rajapakse and Zhou, 2007; Li *et al.*, 2008; Ramsey *et al.*, 2010). We note that the subgroup DAGs that summarise cluster-specific graphical structure may be useful as summary statistics for the purposes of dimensionality reduction.

The running time of the JEMN

The MDM scores (LPL) have to be obtained before applying the JEMN, and the duration of this step was presented in Section 3.3.3. For this application that consists of 10 nodes and 1200 time points, it took around 18 hours to obtain the scores per dataset, on a 2.7 GHz quad-core Intel Core i7 linux host with 16 GB, considering 2^{n-1} possible sets of

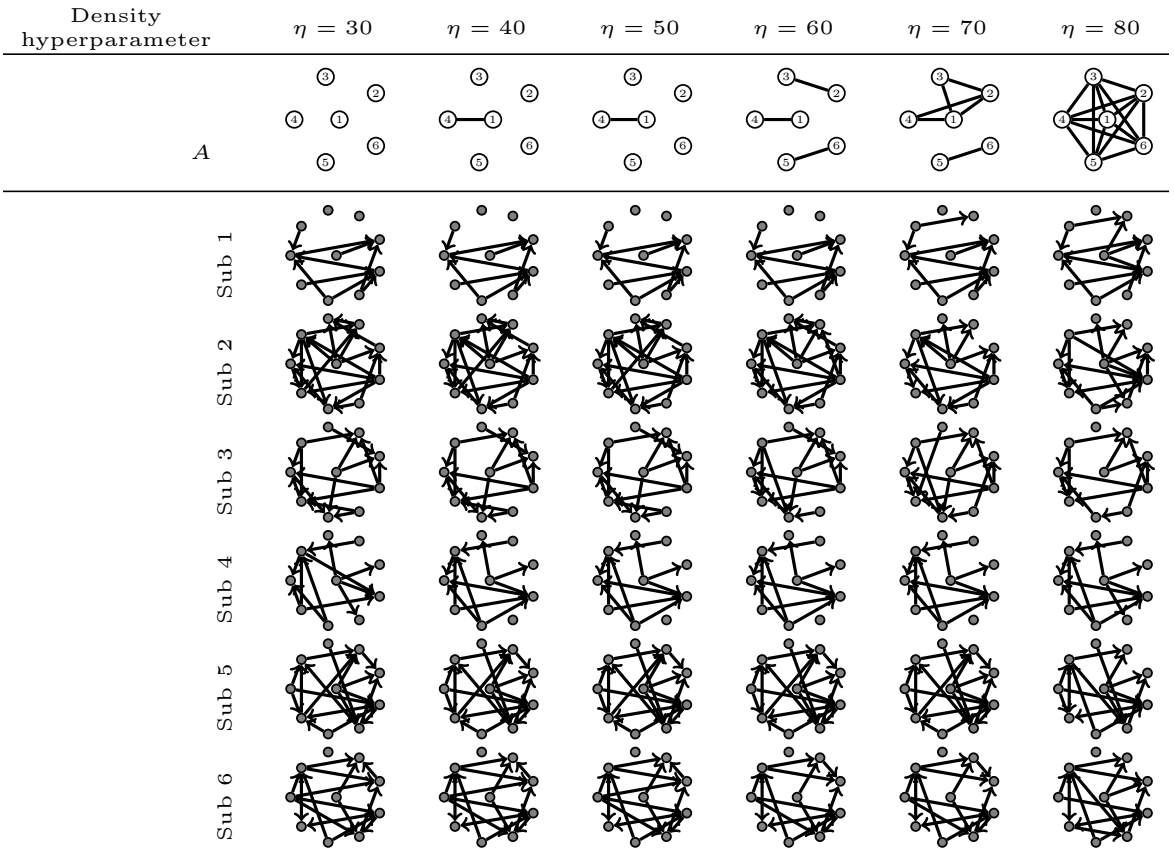


Figure 6.18: Learning multiple DAGs without an exchangeability assumption. Here we simultaneously estimate both subject-specific DAGs and the network A that relates subjects. The regularity hyperparameter $\lambda = 4$ was fixed whilst the density hyperparameter η was varied. Figure 6.13 (b) provides a key.

parents per node, the discount factor chosen in the range from 0.5 to 1.0 with an increment of 0.01, and using the software *R*.

At present an analysis involving $S \leq 10$ subjects, DAGs of size $n \leq 10$ and an in-degree restriction $d_{\max} = 3$ requires approximately 10 minutes of serial computation. Our ongoing research focuses on reducing this computational burden so that exact estimation becomes feasible for much larger datasets. Recent advances in estimation of single DAGs involving thousands of nodes suggests that much progress can be made in this direction (Bartlett and Cussens, 2013; Sheehan *et al.*, 2014).

6.5 Discussion

We developed here the IEMN and the MEMN methods based on the GS approach. We showed that these two approached provide similar results when the hyperparameter λ is

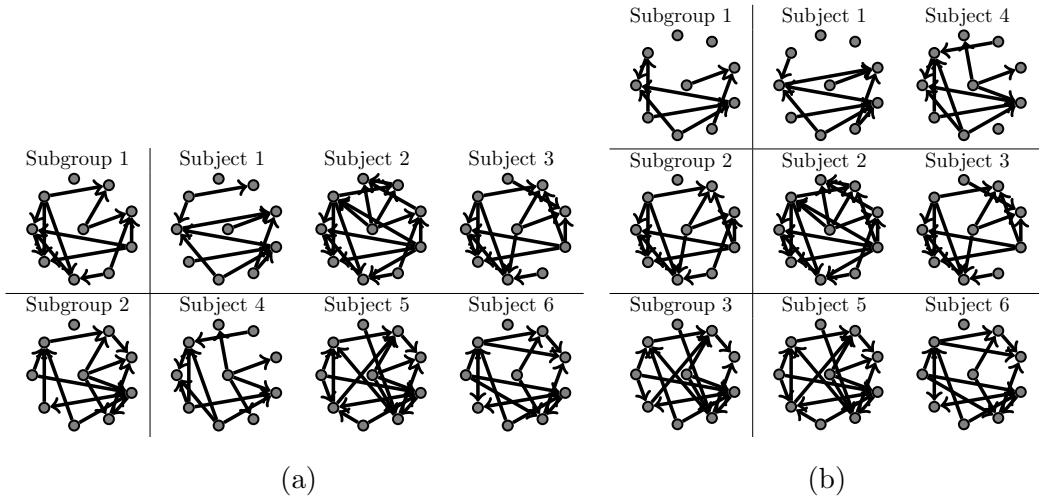


Figure 6.19: Learning networks using the JEMN with (a) G = 2 clusters and (b) G = 3 clusters. We simultaneously estimate subject-specific DAGs, their cluster assignments and the subgroups that summarise graphical structure within each cluster. The regularity hyperparameter was fixed at $\lambda = 4$ and Figure 6.13 (b) provides a key.

small for individual networks and large for group networks. Moreover the higher λ, the denser the individual networks are, but the sparser the group networks are. In this chapter, we discussed some procedures that can be applied in order to estimate λ. The results found here suggest that the estimation of λ depends on the aim of the study, *e.g.*, if one wishes to predict the connectivity for a new subject, then cross-validation can be used, or yet if the focus is on estimating individual networks, then λ can be chosen maximising the posterior distribution, $p(M_i|\mathbf{y})$. In general, the appropriate choice of λ is also related to how homogeneous the group is, and so how much of the information of other datasets should also be included in the analysis of one dataset.

It is not possible to improve the estimation of a particular individual network using the information of other subjects in the IEMN. The MEMN can address this, but may provided denser graphs than the IEMN, as shown above. An exact algorithm that facilitates the joint estimation of multiple DAGs was recently developed in Oates *et al.* (2014), viewing the estimation problem within a hierarchical Bayesian framework and applying advanced techniques from integer linear programming to obtain a *maximum a posteriori* estimate of all DAGs simultaneously. The availability of exact algorithms opens up the opportunity to analyse multi-subject neural connectivity using causal DAG models, whilst leveraging the similarity between subjects in order to improve statistical efficiency and robustness. In Section 6.4.2 we illustrated the scope and applicability of these exact algorithms within

neuroscience, using a small fMRI time course dataset obtained on six subjects, coupled with the MDM-IPA. It is envisaged that exact algorithms will play an important rôle in future studies of neural connectivity.

We can also extend the JEMN approach to estimate directed graphs rather than DAGs. That is, the MDM-DGM can be applied, maximising the scores provided above for each node r independently of other nodes. As shown above, the MDM-DGM usually provides denser graphs than the MDM-IPA. It is expected because the space of possible directed graphs is higher than DAGs for the same number of nodes. Therefore, the use of a penalty function may help identify sparser directed graphs.

Chapter 7

Further Research

In this chapter, we present some ideas about the extensions of the methodologies shown in this thesis, as future work. In Section 7.1, we discuss the use of the non-local priors in the learning network process, whilst, in Section 7.2, we suggest the use of random effect models to estimate connectivity strengths.

7.1 Search methods for the MDM using non-local priors

We have discussed in this thesis the problem that some denser graphs are chosen in model selection processes, albeit with some connectivity strengths close to zero when the sparse graphs hold. To address this, Consonni and La Rocca (2010) developed a novel method to compare pairwise nested models using the fractional Bayes factors (FBF) and moment priors. The FBF was defined by O'Hagan (1995) comparing model M_1 against model M_0 as

$$\begin{aligned} \text{FBF}_{10}(\mathbf{y}; b) &= \frac{w_1(\mathbf{y}|b)}{w_0(\mathbf{y}|b)}, \text{ where} \\ w_k(\mathbf{y}|b) &= \frac{\int f_k(\mathbf{y}|\pi_k) p_k(\pi_k) d\pi_k}{\int f_k^b(\mathbf{y}|\pi_k) p_k(\pi_k) d\pi_k}; \end{aligned} \tag{7.1}$$

$\mathbf{y} = (y(1), \ldots, y(n))$ is the observed sample; $f_k(\mathbf{y}|\pi_k)$ is the sampling density whilst $f_k^b(\mathbf{y}|\pi_k)$ is the likelihood raised to the b-th power; the value of b is fixed depending on the sample size n and assuming $0 < b < 1$; π_k is the parameter of the model M_k; and $p_k(\pi_k)$ is the prior distribution, for $k = 0, 1$.

Consonni and La Rocca (2010) suggested the moment priors for parameter π considering a Gaussian distribution in the context of DAGs, when M_0 is nested to M_1. That

is,

$$p_k(\pi_k) = \prod_{r=1}^{n} \left[V_k(r) \prod_{l \in L_r} \left(\theta_k^{(l)}(r) \right)^{2h} \right], \tag{7.2}$$

where $\pi_k = \bigcup_{r=1}^{n} (V_k(r), \boldsymbol{\theta}_k(r))$, being $V_k(r)$ the conditional variance and $\boldsymbol{\theta}_k(r) = (\theta_k^{(1)}(r), \ldots, \theta_k^{(p_r)}(r))$ the regression parameters for node r and model M_k; and L_r is the subset of the parents of node r in M_1 but not in M_0. The hyperparameter h sets zero for M_0, returning the local prior, whilst h is defined as a positive integer number for M_1 (usually assuming the value 1). Note that the values of $\theta_1^{(l)}(r)$ near zero decrease the evidence for the largest model.

Therefore, based on these moment priors given in equation (7.2), Consonni and La Rocca (2010) provided the exactly formula for $w_k(y(r)|\mathbf{x}(r), b)$ in equation (7.1), for each node r, where $\mathbf{x}(r) = (y(1), \ldots, y(r-1))$ for $r > 1$ and $\mathbf{x}(r) = \emptyset$ for $r = 1$. Then,

$$\begin{aligned} \mathrm{FBF}_{10}(\mathbf{y}; b) &= \frac{w_1(\mathbf{y}|b)}{w_0(\mathbf{y}|b)} \\ &= \prod_{r=1}^{n} \frac{w_1(\mathbf{y}(r)|\mathbf{x}(r), b)}{w_0(\mathbf{y}(r)|\mathbf{x}(r), b)} \\ &= \prod_{r=1}^{n} \mathrm{FBF}_{(10)}^{(r)}(\mathbf{y}(r); \mathbf{x}(r), b). \end{aligned}$$

It is notable that the modularity property is also applied here so that it is only necessary to evaluate $\mathrm{FBF}_{(10)}^{(r)}(\mathbf{y}(r); \mathbf{x}(r), b)$ for nodes in which the number of their parents increases from model M_0 to M_1. Therefore, we expect to develop the ideas given here in the context of searching MDMs, and also to assess the inclusion (or exclusion) of a set of nodes as parents of a particular node, as discussed in parent-child monitor (Section 3.5.2).

7.2 The Multiregression Dynamic Hierarchical Models

In the two previous chapters, we provided some possibilities of pooling information from different datasets into a unique analysis in order to obtain more precise and robust estimates. However, the group analysis discussed in the literature (see Chapters 5 and 6) is mainly focused on estimating the graphical structures, and so the connectivity strengths are usually found as the average of estimated parameters over datasets. For future work, we plan to extend the group analysis methods discussed before to incorporate random effect models to

estimate the connectivity strengths. The idea is (i) firstly the homogeneous subgroups are defined using the pairwise logBF separation and cluster analysis, as presented in Chapter 5; (ii) then one group analysis method provided in Chapter 6, *e.g.* the JEMN, can be applied to estimate the graphical structure for every subgroup (or the entire group if it is homogeneous); (iii) finally, the multiregression dynamic hierarchical model (MDHM) is fitted to estimate connectivity strengths per subgroup, considering two levels: one for brain regions and other for subjects (and it is also possible to add one more level for replications).

We plan to develop the MDHM based on the class of *dynamic hierarchical model* which consists of four parts: the observation equation, the structural equations, the system equation and initial information (Gamerman and Migon, 1993). The observation equation, the system equation and initial information are defined similarly to the MDM. However, the structural equations specify the structure of parameters hierarchy, as shown below.

The *observation equation* can be defined as

$$\mathbf{Y}_t = \mathbf{F}_{1t}'\boldsymbol{\theta}_{1t} + \mathbf{v}_{1t}, \qquad \mathbf{v}_{1t} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_{1t});$$

where $\mathbf{Y}_t$ is a observed vector with dimension n at time t; $t = 1, \ldots, T$; and $\mathbf{F}_{1t}$ is a known covariate matrix. The p_1-dimensional time-varying regression coefficient $\boldsymbol{\theta}_{1t}$ represents the effects of covariates into the observed time series.

The *structural equations* are written as

$$\begin{aligned} \boldsymbol{\theta}_{1t} &= \mathbf{F}_{2t}'\boldsymbol{\theta}_{2t} + \mathbf{v}_{2t}, \qquad \mathbf{v}_{2t} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_{2t}); \\ &\vdots \\ \boldsymbol{\theta}_{kt} &= \mathbf{F}_{kt}'\boldsymbol{\theta}_{kt} + \mathbf{v}_{kt}, \qquad \mathbf{v}_{kt} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_{kt}); \end{aligned}$$

where $\mathbf{F}_{it}$ is a known covariate matrix at the level i; $i = 1, \ldots, k$. The p_i-dimensional time-varying regression coefficient at level i is $\boldsymbol{\theta}_{it}$, satisfying $p_1 > p_2 > \ldots > p_k$.

The *system equation* is set as

$$\boldsymbol{\theta}_{kt} = \mathbf{G}_t\boldsymbol{\theta}_{k,t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t).$$

Thus, only the regression parameters of the last level k are allowed to evolve in time according

to a known matrix $\mathbf{G}_t$. All residuals $\mathbf{v}_{1t}, \ldots, \mathbf{v}_{kt}$ and $\mathbf{w}_t$ are assumed independent with known variance-covariance matrix $\mathbf{V}_{1t}, \ldots, \mathbf{V}_{kt}$ and $\mathbf{W}_t$, respectively.

Finally the *initial information* is written as

$$(\boldsymbol{\theta}_{i0}|y_0) \sim \mathcal{N}(\mathbf{m}_{i0}, \mathbf{C}_{i0});$$

where $i = 1, \ldots, k$; the mean vector $\mathbf{m}_{i0}$ with dimension p_i is an initial estimate of the regression parameters and $\mathbf{C}_{i0}$ is the $p_i \times p_i$ variance-covariance matrix.

When the observational variances are unknown, we can assume $\mathbf{V}_{1t} = \sigma^2\mathbf{I}_n$, and reparameterise the model as before, *i.e.* $\mathbf{V}_{it} = \sigma^2\mathbf{V}^*_{it}$, for $i > 1$, $\mathbf{W}_t = \sigma^2\mathbf{W}^*_t$, and $\mathbf{C}_{i0} = \sigma^2\mathbf{C}^*_{i0}$. By defining $\phi^{-1} = \sigma^2$, we can include this prior in the model:

$$(\phi|y_0) \sim \mathcal{G}\left(\frac{n_0}{2}, \frac{d_0}{2}\right).$$

The posterior distributions of parameters ($\boldsymbol{\theta}$'s and ϕ) and the predictive distributions can be found through the Kalman filtering algorithm (see Gamerman and Migon, 1993, for details).

Bibliography

Agresti A. (2002). *Categorical data analysis*. John Wiley, New York.

Ali R.A., Richardson T.S., Spirtes P. (2009). Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37, 2808-2837.

Allen E.A., Damaraju E., Plis S.M., Erhardt E.B., Eichele T., Calhoun V.D. (2012). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, bhs352.

Anacleto Junior O., Queen C., Albers C. (2013a). Multivariate forecasting of road traffic flows in the presence of heteroscedasticity and measurement errors. *Journal of the Royal Statistical Society: Series C* (Applied Statistics), 62(2), 251-270.

Anacleto Junior O., Queen C., Albers C. (2013b). Forecasting multivariate road traffic flows using Bayesian dynamic graphical models, splines and other traffic variables. *Australian & New Zealand Journal of Statistics*, 55(2), 69-86.

Arnhold J., Grassberger P., Lehnertz K., Elger C.E. (1999). A robust method for detecting interdependences: application to intracranially recorded EEG. *Physica D: Nonlinear Phenomena*, 134(4), 419-430.

Ashburner J., Friston K.J. (2007). Non-linear registration. In: Friston K.J., Ashburner J., Kiebel S., Nichols T.E., Penny W.D. (Eds) *Statistical parametric mapping: the analysis of functional brain images*, 1st ed, London: Elsevier/Academic Press.

Aston J.A.D., Kirch C. (2012). Evaluating stationarity via change-point alternatives with applications to fMRI data. *The Annals of Applied Statistics*, 6(4), 1906-1948.

Baba K., Shibata R., Sibuya M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics*, 46(4), 657-664.

Banerjee O., El Ghaoui L., d'Aspremont A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.* 9, 485-516.

Banerjee O., Ghaoui L.E., d'Aspremont A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*, 9, 485-516.

Bartlett M., Cussen J. (2013). Advances in Bayesian network learning using integer programming. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*. AUAI Press. To appear.

Beckmann C.F., Smith S.M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *Trans. Med. Imaging*, 23(2), 137-152.

Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B 57 (1), 289-300.

Bernardo J. M. (1985). Discussion of paper by A. F. M. Smith and L. I. Pettit. *In Bayesian Statistics 2*, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 492-3. Amsterdam: North- Holland; and Valencia University Press.

Bhattacharya S., Maitra R. (2011). A nonstationary nonparametric Bayesian approach to dynamically modelling effective connectivity in functional magnetic resonance imaging experiments. *The Annals of Applied Statistics*, 5 (2B), 1183-1206.

Bhattacharya S., Ringo Ho M., Purkayastha S. (2006). A Bayesian approach to modelling dynamic effective connectivity with fMRI data. *NeuroImage*, 30, 794-812.

Binder J.R., Frost J.A., Hammeke T.A., Bellgowan P.S., Rao S.M., Cox R.W. (1999). Conceptual processing during the conscious resting state. A functional MRI study. *J. Cogn. Neurosci.* 11, 80-95.

Biswal B., Yetkin F.Z., Haughton V.M., Hyde J.S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson Med.* 34, 537-541.

Bollen K.A., and Long S.J. (1993). *Testing Structural Equation Models*. SAGE Focus Edition, vol. 154, ISBN 0-8039-4507-8.

Box G., Jenkins G., Reinsel G. (1994). *Time series analysis: forecasting and control*, Oakland, California: Holden-Day.

Bressler S.L., and McIntosh A.R. (2007). The role of neural context in large-scale neurocognitive network operations. In *Handbook of Brain Connectivity*, Jirsa V., McIntosh A.R. (eds.), 403-420, New York: Springer.

Bruce A.G., Martin, R.D. (1989). Leave k-out diagnostics for time series (with discussion). *J. R. Statist. Soc.* B, 51, 363-424.

Buxton R.B., Wong E.C., Frank L.R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the Balloon model. *MRM*, 39, 855-864.

Chang C., Glover G.H. (2010). Time-frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage*, 50, 81-98.

Chang C., Thomason M.E., Glover G.H. (2008). Mapping and correction of vascular hemodynamic latency in the BOLD signal. *NeuroImage*, 43, 90-102.

Chickering M. (2002). Optimal structure identification with greedy search. *Mach. Learn. Res.* 3, 507-554.

Claassen T., and Heskes, T. (2012). A Bayesian Approach to Constraint Based Causal Inference. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 207-216.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1),37-46.

Cohen D. (2007). Review: Facebook for scientists? *Nature Network*. BMJ: British Medical Journal, 335(7616), 401.

Consonni G., and La Rocca, L. (2010). Moment Priors for Bayesian Model Choice with Applications to Directed Acyclic Graphs. *Bayesian Statistics*, 9(9), 119-144.

Cordes D., Haughton V.M., Arfanakis K., Wendt G.J., Turski P.A., Moritz C.H., Quigley M.A., Meyerand M.E. (2000). Mapping functionally related regions of brain with functional connectivity MR imaging. *Am. J. Neuroradiol.* 21(9), 1636-1644.

Cover T. M., Thomas J.A. (2000). *Elements of Information Theory.* Wiley. New York.

Cowell R.G. (2013). A simple greedy algorithm for reconstructing pedigrees. *Theoretical Population Biology*, 83, 55-63.

Cowell R.G., Dawid A.P., Lauritzen S.L., Spiegelhalter D.J. (1999). *Probabilistic Networks and Expert Systems.* Springer-Verlag, New York.

Cribben I., Haraldsdottir R., Atlas L.Y., Wager T.D., Lindquist M.A. (2012). Dynamic connectivity regression: determining state-related changes in brain connectivity. *NeuroImage*, 61(4), 907-920.

Cussens J. (2010). Maximum likelihood pedigree reconstruction using integer programming. In *Proceedings of the Workshop on Constraint Based Methods for Bioinformatics (WCB-10)*, Edinburgh, July.

Cussens J. (2011). Bayesian network learning with cutting planes. In Fabio G. Cozman and Avi Pfeffer, editors, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 153-160, Barcelona. AUAI Press.

Danaher P., Wang P., Witten D.M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc.* B, 76(2), 373-397.

Daunizeau J., Friston K.J., Kiebel S.J. (2009). Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D*, 238,2089-2118.

Dauwels J., Vialatte F., Musha T., Cichocki A. (2010). A comparative study of synchrony measures for the early diagnosis of Alzheimer's disease based on EEG. *NeuroImage*, 49(1), 668-693.

David O., Guillemain I., Saillet S., Reyt S., Deransart C., Segebarth C., Depaulis A. (2008). Identifying neural drivers with functional MRI: an electrophysiological validation. *PLoS Biol.* 6, 2683-2697.

Dean T. (1990). Coping with Uncertainty in a Control System for Navigation and Exploration. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, 1010-1015. Menlo Park, Calif.: American Association for Artificial Intelligence.

Dehmer M. (2011). *Structural Analysis of Complex Networks.* 1st Edition, XIV. Birkhäuser Boston.

Denison D.G.T., Holmes C.C., Mallick B.K., Smith A.F.M. (2002). *Bayesian methods for nonlinear classification and regression.* Wiley. Chapter 2.

Ding M., Chen Y., Bressler S.L. (2006). Granger Causality: Basic Theory and Application to Neuroscience, in *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications* (eds B. Schelter, M. Winterhalder and J. Timmer), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

Durbin J., Koopman S.J. (2001). *Time Series Analysis by State Space Methods.* Oxford University Press.

Everitt B., and Hothorn T. (2009). *A Handbook of Statistical Analyses Using R.* 2nd ed. Chapman & Hall/CRC.

Everitt B., Landau S., Leese M., Stahl D. (2011). *Cluster analysis.* 5th ed. Chichester: Wiley.

Finger S. (1994). *Origins of neuroscience: A history of explorations into brain function.* New York: Oxford University Press.

Fox M.D., Snyder A.Z., Vincent J.L., Corbetta M., Van Essen D.C., Raichle M.E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci.* 102 (27), 9673-9678.

Friedman J., Hastie T., Tibshirani R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3), 432-441.

Friston K.J. (2005). Models of brain function in neuroimaging. *Annu Rev Psychol*, 56, 57-87.

Friston K.J. (2011). Functional and Effective Connectivity: a review. *Brain Connectivity*, 1(1), 13-36.

Friston K.J., Harrison L., Penny W. (2003). Dynamic causal modelling. *NeuroImage*, 19, 1273-1302.

Friston K.J., Mechelli A., Turner R., Price C.J. (2000). Nonlinear responses in fMRI: the Balloon model, Volterra kernels and other hemodynamics. *NeuroImage*, 12, 466-477.

Fruhwirth-Schnatter S. (1995). Bayesian Model Discrimination and Bayes Factor for Linear Gaussian State Space Models. *Journal of the Royal Statistical Society*, Series B, 57, 237-246.

Fruhwirth-Schnatter S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Gamerman D. (1997). *Markov Chain Monte Carlo: stochastic simulation for bayesian inference*. 1.ed. London: Chapman & Hall, 192-211.

Gamerman D., Migon H. (1993). Dynamic hierarchical models. *J. Roy. Statist. Soc.* B 55(3), 629-642.

Gates K.M. (2012). Identifying subgroups using fMRI connectivity maps. Paper presented at the annual meeting for the Society for Neuroscience, New Orleans.

Gates K.M., Molenaar P.C.M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, 63, 310-319.

Ge T., Kendrick K.M., Feng J. (2009). A novel extended Granger causal model approach demonstrates brain hemispheric differences during face recognition learning. *PLoS Comput. Biol.* 5.

Gibbens R. (2000). Control and Pricing for Communication Networks. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, Vol. 358, No. 1765, Science into the Next Millenium: Young Scientists Give Their Visions of the Future: II. Mathematics, Physics and Engineering, 331-341.

Goldenberg A., Zheng A.X., Fienberg S.E., Airoldi E.M. (2009). A Survey of Statistical Networks Models. *Foundations & Trends in Machine Learning*, in press.

Goldman R.P. (1990). *A Probabilistic Approach to Language Understanding*. PhD thesis, Department of Computer Science, Brown University, Providence, RI Technical Report CS-90-34.

Gonçalves M.S., Hall D.A., Johnsrude I.S., Haggard M.P. (2001). Can meaningful effective connectivities be obtained between auditory cortical regions? *NeuroImage*, 14, 1353-1360.

Granger C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424-438.

Greicius M.D., Krasnow B., Reiss A.L., Menon V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 100, 253-258.

Harrison J., West M. (1991). Dynamic linear-model diagnostics. *Biometrika*, 78(4), 797-808.

Havlicek M., Jan J., Brazdil M., Calhoun V.D. (2010). Dynamic Granger causality based on Kalman filter for evaluation of functional network connectivity in fMRI data. *NeuroImage*, 53, 65-77.

Hayasaka S., & Laurienti P.J. (2010). Comparison of characteristics between region-and voxel-based network analyses in resting-state fMRI data. *NeuroImage*, 50(2), 499-508.

Heard N.A., Holmes C.C., Stephens D.A. (2006). A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. *Journal of the American Statistical Association*, 101(473), 18-29.

Heckerman D. (1990). Similarity networks for the construction of multiple-fault belief networks. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, Boston, MA, 32-39.

Heckerman D. (1999). A Tutorial on Learning with Bayesian Networks. In *Learning in Graphical Models*, M. Jordan, ed. MIT Press, Cambridge, MA.

Hill S.M., Lu Y., Molina J., Heiser L.M., Spellman P.T., Speed T.P., ..., Mukherjee S. (2012). Bayesian Inference of Signaling Network Topology in a Cancer Cell Line. *Bioinformatics*, 28(21), 2804-2810.

Honey C.J., Sporns 0., Cammoun L., Gigandet X., Thiran J.P., Meuli R., Hagmann P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc Natl Acad Sci USA*, 106, 2035-2040.

Hotelling H. (1953). New light on the correlation coefficient and its transform. *Journal of the Royal Statistical Society*, Series B15, 193-232.

Hoyle R.H. (1995). *Structural Equation Modeling: Concepts, Issues, and Applications*. SAGE, ISBN 0-8039-5318-6.

Hyvärinen A., Oja E. (2000). Independent Component Analysis: Algorithms and Application. *Neural Networks*, 13(4-5), 411-430.

Jaakkola T., Sontag D., Globerson A., Meila M. (2010). Learning Bayesian network structure using LP relaxations. In *Proceedings of 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, 9, 358-365. Journal of Machine Learning Research Workshop and Conference Proceedings.

Jeffreys H. (1961). *Theory of Probability* (3rd ed.). Oxford University Press, London.

Jenkinson M., Bannister P., Brady J. M., Smith, S. M. (2002). Improved Optimisation for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2), 825-841.

Johnson W., Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *J. Am. Statist. Assoc.* 78, 137-44.

Kalisch M., Bühlmann P. (2008). Robustification of the PC-Algorithm for Directed Acyclic Graphs. *Journal of Computational and Graphical Statistics*, 17(4).

Kherif F., Poline J.-B., Mériaux S., Benali H., Flandin G., Brett M. (2004). Group analysis in functional neuroimaging: selecting subjects using similarity measures. *NeuroImage*, 20(4), 2197-2208.

Kim P.M., Lu L.J., Xia Y., Gerstein M.B. (2006). Relating Three-Dimensional Structures to Protein Networks Provides Evolutionary Insights. *Science*, New Series, 314(5807), 1938-1941.

Kiviniemi V., Kantola J.H., Jauhiainen J., Hyvarinen A., Tervonen O. (2003). Independent component analysis of nondeterministic fMRI signal sources. *NeuroImage*, 19 (2), 253-260.

Kline R.B. (2010). *Principles and Practice of Structural Equation Modeling* (3rd Edition). The Guilford Press, ISBN 978-1-60623-877-6

Korb K.B., and Nicholson A.E. (2004). *Bayesian Artificial Intelligence*. Computer Science and Data Analysis. Chapman & Hall / CRC, Boca Raton.

Koster J.T. (1996). Markov properties of nonrecursive causal models. *The Annals of Statistics*, 2148-2177.

Langfelder P., Zhang B., Horvath S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24(5), 719-720.

Lauritzen S.L. (1996). *Graphical Models*. Oxford, United Kingdom: Clarendon Press.

Le Q.A., Doctor J.N. (2011). Probabilistic mapping of descriptive health status responses onto health state utilities using Bayesian networks: an empirical analysis converting SF-12 into EQ-5D utility index in a national US sample. *Med Care*. May, 49(5), 451-60.

Lenartowicz A, McIntosh AR. (2005). The role of anterior cingulate cortex in working memory is shaped by functional connectivity. *J Cogn Neurosci.*, 17(7), 1026-1042.

Leonardi N., Richiardi J., Gschwind M., Simioni S., Annoni J.M., Schluep M., Vuilleumier P., Van De Ville D. (2013). Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. *NeuroImage*, 83, 937-950.

Li B., Daunizeau J., Stephan K.E., Penny W., Hu D., Friston K. (2011). Generalised filtering and stochastic DCM for fMRI. *NeuroImage*. 58, 442-457.

Li J., Wang Z.J., Palmer S.J., McKeown M.J. (2008). Dynamic Bayesian network modeling of fMRI: a comparison of group-analysis methods. *NeuroImage*, Jun, 41(2), 398-407.

Mandeville J.B., Marota J.J., Ayata C., Zararchuk G., Moskowitz M.A., Rosen B., Weisskoff R.M. (1999). Evidence of a cerebrovascular postarteriole Windkessel with delayed compliance. *J. Cereb. Blood Flow Metab*, 19, 679-689.

Mardia K.V., Kent J.T., Bibby J.M. (1979). *Multivariate Analysis*. London, UK: Academic Press.

Marrelec G., Krainik A., Duffau H., Pélégrini-Issac M., Lehéricy S., Doyon J., Benali H. (2006). Partial correlation for functional brain interactivity investigation in functional MRI. *NeuroImage*, 32, 228-237.

Mazoyer B., Zago L., Mellet E., Bricogne S., Etard O., Houdé O., . . ., Tzourio-Mazoyer N. (2001). Cortical networks for working memory and executive functions sustain the conscious resting state in man. *Brain Res. Bull.* 54 (3), 287-298.

McGonigle D.J., Howseman A.M., Athwal B.S., Friston K.J., Frackowiak R.S., Holmes A.P. (2000). Variability in fMRI: an examination of intersession differences. *NeuroImage*, 11, 708-734.

McIntosh AR. (2000). Towards a network theory of cognition. *Neural Network*, 13(8-9), 861-70.

McKeown M.J., Makeig S., Brown G.G., Jung T.P., Kindermann S.S., Bell A.J., Sejnowski T.J. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Hum. Brain Mapping*. 6 (3), 160-188.

Mechelli A., Penny W.D., Price C.J., Gitelman D.R., Friston K.J. (2002). Effective connectivity and intersubject variability: using a multisubject network to test differences and commonalities. *NeuroImage*, 17 (3), 1459-1469.

Meek C. (1995). Causal inference and causal explanation with background knowledge. *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, 403-410.

Meek C. (1997). *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University.

Mohan K., Chung M., Han S., Fazel M., Witten D., Lee S. (2012). Structured sparse learning of multiple Gaussian graphical models. *Proceedings of the 26th Conference on Advances in Neural Information Processing Systems*, 620-628.

Oates C.J. (2013). *Bayesian Inference for Protein Signalling Networks*. PhD Thesis. The University of Warwick: U.K. Chapter 4.

Oates C.J., Smith J.Q., Mukherjee S., Cussens J. (2014). Exact Estimation of Multiple Directed Acyclic Graphs. *CRiSM Working Paper, University of Warwick*, 14(07).

O'Hagan A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society*. Series B (Methodological), 57 (1), 99-138.

Patel R., Bowman F., Rilling J. (2006). A Bayesian approach to determining connectivity of the human brain. *Hum. Brain Mapping*. 27, 267-276.

Pearl J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Penfold C.A., Buchanan-Wollaston V., Denby K.J., Wild D.L. (2012). Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics*, 28(12), i233-i241.

Penny W., Ghahramani Z., Friston K. (2005). Bilinear dynamical systems. *Phil. Trans. R. Soc. B*, 360, 983-993.

Pereda E., Quiroga R., Bhattacharya J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Prog. Neurobiol.* 77 (1-2), 1-37.

Petris G., Petrone S., Campagnoli P. (2009). *Dynamic Linear Models with R*. Springer, New York.

Poldrack R.A., Mumford J.A., Nichols T.E. (2011). *Handbook of fMRI Data Analysis*. Cambridge University Press.

Queen C.M., Wright, B.J., Albers C.J. (2008). Forecast covariances in the linear multiregression dynamic model. *J. Forecast.*, 27, 175-191.

Queen C.M., and Albers C.J. (2009). Intervention and causality: Forecasting traffic flows using a dynamic bayesian network. *Journal of the American Statistical Association*. June. 104(486), 669-681.

Queen C.M., and Smith J.Q. (1993). Multiregression dynamic models. *Journal of the Royal Statistical Society*, Series B, 55, 849-870.

Quian Quiroga R., Arnhold J., Grassberger P. (2002). Learning driver-response relationships from synchronization patterns. *Phys. Rev. E*, 61, 5142.

Quian Quiroga R., Kraskov A., Kreuz T., Grassberger P. (2002). Performance of different synchronization measures in real data: a case study on electroencephalographic signals. *Phys. Rev. E*, 65 (4), 41903.

Raichle M.E. (2010). Two views of brain function. *Trends in Cognitive Sciences*, 14 (4), 180-190.

Raichle M.E., MacLeod A.M., Snyder A.Z., Powers W.J., Gusnard D.A., Shulman G.L. (2001). A default mode of brain function. *Proc.Natl. Acad. Sci. U. S. A.* 98, 676-682.

Rajapakse J.C., and Zhou, J. (2007). Learning effective brain connectivity with dynamic Bayesian networks. *NeuroImage*, 37, 749-760.

Ramsey J.D., Hanson S.J., Hanson C., Halchenko Y.O., Poldrack R.A., and Glymour C. (2010). Six Problems for Causal Inference from fMRI. *NeuroImage*, 49, 1545-1558.

Richardson T. (1996). A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models. *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, 462-469, August 01-04, Portland, OR.

Ringach D.L. (2009). Spontaneous and driven cortical activity: implications for computation. *Curr. Opin. Neurobiol.* 19, 439-444.

Robinson L.F., Wager T.D., Lindquist M.A. (2010). Change point estimation in multi-subject fMRI studies. *NeuroImage*, 49(2), 1581-92.

Roebroeck A., Formisano E., Goebel R. (2011). The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage*, 58, 296-302.

Ryali S., Supekar K., Chen T., Menon V. (2011). Multivariate dynamical systems models for estimating causal interactions in fMRI. *NeuroImage*, 54, 807-823.

Salimi-Khorshidi G., Douaud G., Beckmann C.F., Glasser M.F., Griffanti L., Smith S.M. (2014). Automatic Denoising of Functional MRI Data: Combining Independent Component Analysis and Hierarchical Fusion of Classifiers. *NeuroImage*, 90, 449-468.

Sampson F.S. (1968). *A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships*. PhD thesis, Cornell University.

Santos A. (2014). *Dynamic Bayesian smooth transition autoregressive models.* PhD thesis, The Open University.

Schwarz G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6, 461-464.

Scott J.G., and Berger J.O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *Ann. Stat.* 38(5), 2587-2619.

Sheehan *et al.* (2014). Maximum Likelihood Reconstruction of Very Large Pedigrees. *Theoretical Population Biology*, in submission.

Shimizu S., Hoyer P.O., Hyvärinen A., Kerminen A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* 7, 2003-2030.

Shulman G.L., Fiez J.A., Corbetta M., Buckner R.L., Miezin F.M., Raichle M.E., Petersen S.E. (1997). Common blood flow changes across visual tasks: II. Decreases in cerebral cortex. *J. Cogn. Neurosci.* 9, 648-663.

Smith A.F.M., Pettit, L.I. (1985). Outliers and influential observations in linear models. *In Bayesian Statistics* 2, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 473-94. Amsterdam: North-Holland; and Valencia University Press.

Smith J.F., Pillai A., Chen K., Horwitz B. (2010). Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems. *NeuroImage*, 52, 1027-1040.

Smith J.F., Pillai A., Chen K., Horwitz B. (2011). Effective connectivity modeling for fMRI: six issues and possible solutions using linear dynamic systems. *Front. Syst. Neurosci.* 5(104).

Smith J.Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, 4, 283-291.

Smith J.Q. and Croft J. (2003). Bayesian networks for discrete multivariate data: an algebraic approach to inference. *J. of Multivariate Analysis*, 84, 387-402.

Smith S.M., Andersson J., Auerbach E.J., Beckmann C., Bijsterbosch J., Douaud G., ..., Glasser M.F. (2013). Resting-state fMRI in the Human Connectome Project. *NeuroImage*, 80, 144-168.

Smith S.M., Bandettini P.A., Miller K.L., Behrens T.E.J., Friston K.J., David O., ..., Nichols T.E. (2012). The danger of systematic bias in group-level FMRI-lag-based causality estimation. *NeuroImage*, 59(2), 1228-9.

Smith S.M., Fox P.T., Miller K.L., Glahn D.C., Fox P.M., Mackay C.E., ..., Beckmann C.F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U. S. A.* 106, 13040-13045.

Smith, S.M., Miller K.L., Salimi-Khorshidi G., Webster M., Beckmann C., Nichols T., Ramsey J., Woolrich M. (2011). Network modeling methods for FMRI. *NeuroImage*, 54(2), 875-891.

Spirtes P. (1995). Directed Cyclic Graphical Representation of Feedback Models. *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. by Philippe Besnard and Steve Hanks, Morgan Kaufmann Publishers, Inc., San Mateo.

Spirtes P., Glymour C.N., Scheines R. (2000). *Causation, prediction, and search*, 2nd ed. Cambridge, Mass.: MIT Press.

Sporns O. (2011). *Networks of the Brain*, MIT Press.

Sporns O. (2013). Network attributes for segregation and integration in the human brain. *Current Opinion in Neurobiology*, 23, 162-171.

Stein J.L., Wiedholz L.M., Bassett D.S., Weinberger D.R., Zink C.F., Mattay V.S., Meyer-Lindenberg A. (2007). A validated network of effective amygdala connectivity. *NeuroImage*, 36(3), 736-745.

Steinsky B. (2003). Enumeration of labeled chain graphs and labeled essential directed acyclic graphs. *Discrete Mathematics*, 270, 267-278.

Stephan K.E., Kasper L., Harrison L.M., Daunizeau J., den Ouden H.E., Breakspear M., Friston K.J. (2008). Nonlinear dynamic causal models for fMRI. *NeuroImage*, 42, 649-662.

Stephan K.E., Penny W.D., Moran R.J., den Ouden H.E.M., Daunizeau J., Friston K.J. (2010). Ten simple rules for dynamic causal modeling. *NeuroImage*, 49(4), 3099-3109.

Sucar L.E. (2006). *Introducción a Redes Bayesianas.* Sierra (Ed.), Aprendizaje Automático, Pearson.

Sugihara G., Kaminaga T., Sugishita M. (2006). Interindividual uniformity and variety of the "Writing center": A functional MRI study. *NeuroImage*, 32, 1837-1849.

Sun S., and Zhang C. (2006). A Bayesian Network Approach to Traffic Flow Forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1).

Valdés-Sosa P.A., Roebroeck A., Daunizeau J., Friston K. (2011). Effective connectivity: Influence, causality and biophysical modeling. *NeuroImage*, 58, 339-361.

Van Essen D.C., Smith S.M., Barch D.M., Behrens T.E.J., Yacoub E., Ugurbil K. (2013). The WU-Minn Human Connectome Project: An Overview. *NeuroImage*, 80, 62-79.

Varoquaux G., Sadaghiani S., Pinel P., Kleinschmidt A., Poline J.B., Thirion B. (2010). A group model for stable multi-subject ICA on fMRI datasets. *NeuroImage*, 51(1), 288-99.

West M., and Harrison P.J. (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). New York: Springer-Verlag.

Wiener N. (1956). The theory of prediction. In: *E. F. Beckenbach (Ed) Modern Mathermatics for Engineers*, Chap 8. McGraw-Hill, New York.

Williams H.P. (2009). *Logic and Integer Programming*. Springer.

Yamashita O., Sadato N., Okada T., Ozaki T. (2005). Evaluating frequency-wise directed connectivity of BOLD signals applying relative power contribution with the linear multivariate time-series models. *NeuroImage*, 25, 478-490.

Zhang J., Li X., Li C., Lian Z., Huang X., Zhong G., ... Liu T. (2013). Inferring functional interaction and transition patterns via dynamic bayesian variable partition models. *Human brain mapping*.

Zheng X., and Rajapakse, J.C. (2006). Learning functional structure from fMR images. *NeuroImage*, 31(4), 1601-1613.

Zhou S., Lafferty J., Wasserman L. (2010). Time Varying Undirected Graphs. *Machine Learning*, 80 (2-3), 295-319.

Zou C., and Feng J. (2009). Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics*, 10, 401.

A Supplemental Material for Chapter 3

In this appendix, we provide details about the filtering and smoothing equations based on the Kalman filter, the one-step forecast distribution and the marginal forecast distribution (West and Harrison, 1997; Queen and Smith, 1993; Queen *et al.*, 2008). For simplicity, we are considering here that the set of parents of $Y_t(r)$ are observed in the time t and so they are not explicit in distributions as random variables (except in the calculation of the marginal forecast distribution).

Filtered Distributions

The filtered densities are defined assuming firstly

$$(\boldsymbol{\theta}_{t-1}(r)|\mathbf{y}^{t-1}, \phi(r)) \sim \mathcal{N}\left(\mathbf{m}_{t-1}(r), \mathbf{C}^*_{t-1}(r)\phi^{-1}(r)\right) \text{ and} \tag{7.3}$$

$$(\phi(r)|\mathbf{y}^{t-1}(r)) \sim \mathcal{G}\left(\frac{n_{t-1}(r)}{2}, \frac{d_{t-1}(r)}{2}\right). \tag{7.4}$$

Thus, the marginal distribution of $\boldsymbol{\theta}_{t-1}(r)$ given the past is written as

$$(\boldsymbol{\theta}_{t-1}(r)|\mathbf{y}^{t-1}) \sim \mathcal{T}_{n_{t-1}(r)}\left(\mathbf{m}_{t-1}(r), \mathbf{C}_{t-1}(r)\right), \tag{7.5}$$

a noncentral t distribution with $n_{t-1}(r)$ degrees of freedom and parameters $\mathbf{m}_{t-1}(r)$ and $\mathbf{C}_{t-1}(r) = S_{t-1}(r)\mathbf{C}^*_{t-1}(r)$, where $S_{t-1}(r) = \frac{1}{\mathbb{E}[\phi(r)|\mathbf{y}^{t-1}(r)]} = \frac{d_{t-1}(r)}{n_{t-1}(r)}$.

By equations (3.1) and (7.3),

$$(\boldsymbol{\theta}_t(r)|\mathbf{y}^{t-1}(r), \phi(r)) \sim \mathcal{N}\left(\mathbf{m}_{t-1}(r), \mathbf{R}^*_t(r)\phi^{-1}(r)\right), \tag{7.6}$$

where $\mathbf{R}_t^*(r) = \mathbf{C}_{t-1}^*(r) + \mathbf{W}_t^*(r)$. Thus, from this result and by observation equation,

$$(Y_t(r)|\mathbf{y}^{t-1}(r), \phi(r)) \sim \mathcal{N}\left(f_t(r), Q_t^*(r)\phi^{-1}(r)\right), \tag{7.7}$$

where $f_t(r) = \mathbf{F}_t^{'}(r)\mathbf{m}_{t-1}(r)$ and $Q_t^*(r) = \mathbf{F}_t^{'}(r)\mathbf{R}_t^*(r)\mathbf{F}_t(r) + 1$.

The conditional posterior distribution of $\boldsymbol{\theta}_t(r)$ given $\phi(r)$ is found through the property of multivariate Gaussian distribution. Consider the equations (7.6) and (7.7),

$$(\boldsymbol{\theta}_t(r)|\mathbf{y}^t(r), \phi(r)) \sim \mathcal{N}\left(\mathbf{m}_t(r), \mathbf{C}_t^*(r)\phi^{-1}(r)\right), \tag{7.8}$$

where

$$\begin{aligned}
\mathbf{m}_t(r) &= \mathbf{m}_{t-1}(r) + \mathbf{R}_t^*(r)\mathbf{F}_t(r)(y_t(r) - \mathbf{F}_t^{'}(r)\mathbf{m}_{t-1}(r))/Q_t^*(r); \text{ and} \\
\mathbf{C}_t^*(r) &= \mathbf{R}_t^*(r) - \mathbf{R}_t^*(r)\mathbf{F}_t(r)\mathbf{F}_t^{'}(r)\mathbf{R}_t^*(r)/Q_t^*(r).
\end{aligned}$$

Now, using equations (7.4) and (7.7), the posterior distribution of ϕ is found as follows:

$$\begin{aligned}
p(\phi(r)|\mathbf{y}^t(r)) &\propto p(y_t(r)|\mathbf{y}^{t-1}(r), \phi(r)) \times p(\phi(r)|\mathbf{y}^{t-1}(r)) \\
&\propto \phi(r)^{\frac{(n_{t-1}(r)+1)}{2}-1} \exp\left[-\frac{\phi(r)}{2}\left(\frac{(y_t(r) - f_t(r))^2}{Q_t^*(r)} + d_{t-1}(r)\right)\right].
\end{aligned}$$

Therefore,

$$(\phi(r)|\mathbf{y}^t(r)) \sim \mathcal{G}\left(\frac{n_t(r)}{2}, \frac{d_t(r)}{2}\right), \tag{7.9}$$

where $n_t(r) = n_{t-1}(r) + 1$ and $d_t(r) = d_{t-1}(r) + (y_t(r) - f_t(r))^2/Q_t^*(r)$.

The marginal posterior distribution of $\boldsymbol{\theta}_t(r)$ is found by equations (7.8) and (7.9), *i.e.*

$$(\boldsymbol{\theta}_t(r)|\mathbf{y}^t(r)) \sim \mathcal{T}_{n_t(r)}\left(\mathbf{m}_t(r), \mathbf{C}_t(r)\right),$$

where $\mathbf{C}_t(r) = S_t(r)\mathbf{C}_t^*(r)$.

These results were found assuming the equations (7.3) and (7.4). However, these equations are true for $t = 0$, thus, for other t's, the posterior distribution can be found by using the same arguments.

Smoothed Distributions

The smoothed estimation follows retrospective analysis, starting with $t = T - 1$ and continues until $t = 1$. Firstly, the smoothed distribution of $\boldsymbol{\theta}_t(r)$ given the entire time series and the precision $\phi(r)$, for $t = 1, \ldots, T-1$, is written as:

$$p(\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r)) = \int p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)) p(\boldsymbol{\theta}_{t+1}(r)|\mathbf{y}^T, \phi(r)) d\boldsymbol{\theta}_{t+1}(r). \tag{7.10}$$

Suppose firstly that the second integration term is

$$(\boldsymbol{\theta}_{t+1}(r)|\mathbf{y}^T, \phi(r)) \sim \mathcal{N}\left(\mathbf{sm}_{t+1}(r), \mathbf{sC}^*_{t+1}(r)\phi^{-1}(r)\right). \tag{7.11}$$

Using Bayes' theorem, the first integration term is

$$p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)) = \frac{p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r)) p(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\boldsymbol{\theta}_t(r), \boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r))}{p(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r))}.$$

But, $\mathbf{Y}_{t+1}, \ldots, \mathbf{Y}_T$ are independent of $\boldsymbol{\theta}_t(r)$ given $\boldsymbol{\theta}_{t+1}(r)$ (West and Harrison, 1997) and thus $p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)) = p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r))$. This is a gaussian distribution by equations (7.6) and (7.8) with parameters:

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r)\right] &= \mathbf{m}_t(r) + \mathbf{C}^*_t(r)(\mathbf{R}^*_{t+1}(r))^{-1}(\boldsymbol{\theta}_{t+1}(r) - \mathbf{m}_t(r)); \\ \text{var}\left[\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^t, \phi(r)\right] &= (\mathbf{C}^*_t(r) - \mathbf{C}^*_t(r)(\mathbf{R}^*_{t+1}(r))^{-1}\mathbf{C}^*_t(r))\phi^{-1}(r). \end{aligned} \tag{7.12}$$

Returning to initial problem (equation (7.10)), the required density $p(\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r))$ can be seen as the expectation value of $p(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r))$ (equation (7.12)) with respect to $(\boldsymbol{\theta}_{t+1}(r)|\mathbf{y}^T, \phi(r))$ (equation (7.11)). Therefore, by the properties of the multivariate Gaussian distribution, the conditional distribution of $\boldsymbol{\theta}_t(r)$ given $\mathbf{y}^T$ and ϕ is also gaussian

with the following parameters:

$$\begin{aligned}
\mathbf{sm}_t(r) &= \mathbb{E}\left[\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\right)|\mathbf{y}^T, \phi(r)\right] \\
&= \mathbf{m}_t(r) + \mathbf{C}_t^*(r)(\mathbf{R}_{t+1}^*(r))^{-1}(\mathbf{sm}_{t+1}(r) - \mathbf{m}_t(r)); \\
\mathbf{sC}_t^*(r)\phi^{-1}(r) &= \text{var}\left[\boldsymbol{\theta}_t(r)|\mathbf{y}^T, \phi(r)\right] \\
&= \mathbb{E}\left[\text{var}\left(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\right)|\mathbf{y}^T, \phi(r)\right] + \\
&\quad + \text{var}\left[\mathbb{E}\left(\boldsymbol{\theta}_t(r)|\boldsymbol{\theta}_{t+1}(r), \mathbf{y}^T, \phi(r)\right)|\mathbf{y}^T, \phi(r)\right] \\
&= \left[\mathbf{C}_t^*(r) - \mathbf{C}_t^*(r)(\mathbf{R}_{t+1}^*(r))^{-1}(\mathbf{R}_{t+1}^*(r) - \mathbf{sC}_{t+1}^*(r))(\mathbf{R}_{t+1}^*(r))^{-1}\mathbf{C}_t^*(r)\right]\phi^{-1}(r).
\end{aligned}$$

Moreover, as the conditional distribution of $(\phi(r)|\mathbf{y}^T)$ is given by equation (7.9). Then,

$$(\boldsymbol{\theta}_t(r)|\mathbf{y}^T) \sim \mathcal{T}_{n_T(r)}\left(\mathbf{sm}_t(r), \mathbf{sC}_t(r)\right),$$

where $\mathbf{sC}_t(r) = S_T\mathbf{sC}_t^*(r)$.

The equation (7.11) is true for $t = T - 1$, that is

$$(\boldsymbol{\theta}_T(r)|\mathbf{y}^T, \phi(r)) \sim \mathcal{N}\left(\mathbf{sm}_T(r) = \mathbf{m}_T(r), \mathbf{sC}_{t+1}^*(r)\phi^{-1}(r) = \mathbf{C}_T^*(r)\phi^{-1}(r)\right).$$

Therefore, the distributions of $(\boldsymbol{\theta}_t(r)|\mathbf{y}^T)$ for $t = T-1, T-2, \ldots, 1$ can be computaded by backward procedure.

One-step Forecast Distribution

The one-step conditional forecast distribution can be found through the prior distribution of ϕ given the past (equation (7.4)) and the conditional distribution of Y_t given the past and the precision parameter ϕ (equation (7.7)), *i.e.*:

$$(Y_t(r)|\mathbf{y}^{t-1}, \mathbf{x}_t(r)) \sim \mathcal{T}_{n_{t-1}(r)}\left(f_t(r), Q_t(r)\right),$$

where $Q_t(r) = S_{t-1}(r)Q_t^*(r)$. Note that we included here $\mathbf{x}_t(r)$ for differentiating the forecast conditional distribution shown here from the marginal distribution provided below.

The Marginal Forecast Distribution

The expectation and covariance matrix of the marginal forecast distribution of LMDM are derived here. Firstly, note that the parameters of conditional forecast distribution of $(Y_t(r)|\mathbf{x}_t(r))$ given the past (see equation (3.5)) can be written as

$$\begin{aligned} f_t(r) &= \mathbf{F}_t'(r)\mathbf{m}_{t-1}(r) \\ &= \sum_{i=1}^{r} m_{t-1}^{(i)}(r)F_{tr}(i); \\ Q_t(r) &= \frac{d_{t-1}(r)}{n_{t-1}(r)}[\mathbf{F}_t'(r)\mathbf{R}_t^*(r)\mathbf{F}_t(r)+1] \\ &= \frac{d_{t-1}(r)}{n_{t-1}(r)}\left[\sum_{j=1}^{r}\sum_{k=1}^{r} R_{tr}^*(j,k)F_{tr}(j)F_{tr}(k)+1\right]; \end{aligned}$$

where

$$F_{tr}(i) = \begin{cases} 1 & \text{if } i=1, \\ Y_t(i-1) & \text{if } 2\le i\le r \text{ and } Y_t(i-1)\in Pa(r), \\ 0 & \text{otherwise;} \end{cases} \tag{7.13}$$

$m_{t-1}^{(i)}(r)$ is the i^{th} element of the vector $\mathbf{m}_{t-1}(r)$ and $R_{tr}^*(j,k)$ is the $(j,k)^{th}$ element of matrix $\mathbf{R}_t^*(r)$.

Let $\overline{f_t}(r)=\mathbb{E}[Y_t(r)|\mathbf{y}^{t-1}]$, $\overline{\mathbf{f}}_t^r=(\overline{f_t}(1),\ldots,\overline{f_t}(r))'$ and $\mathbf{\Sigma}_t(r)$ be the forecast covariance matrix of $\{Y_t(1),\ldots,Y_t(r)\}$ such that its $(j,k)^{th}$ element is

$$\{\mathbf{\Sigma}_t(r)\}_{jk}=\sigma_t(j,k)=\text{cov}[Y_t(j),Y_t(k)|\mathbf{y}^{t-1}] \qquad j,k=1,\ldots,r.$$

Suppose initially that $\overline{\mathbf{f}}_t^{r-1}$ and $\mathbf{\Sigma}_t(r-1)$ are known. The expectation of the marginal forecast distribution of $Y_t(r)$ given the past is easily found as:

$$\begin{aligned} \overline{f_t}(r) &= \mathbb{E}[Y_t(r)|\mathbf{y}^{t-1}] \\ &= \mathbb{E}[\mathbb{E}\{Y_t(r)|\mathbf{y}^{t-1},\mathbf{F}_t(r)\}|\mathbf{y}^{t-1}] \\ &= \mathbb{E}\left[\sum_{i=1}^{r} m_{t-1}^{(i)}(r)F_{tr}(i)|\mathbf{y}^{t-1}\right] \\ &= \sum_{i=1}^{r} m_{t-1}^{(i)}(r)\mathbb{E}[F_{tr}(i)|\mathbf{y}^{t-1}], \end{aligned} \tag{7.14}$$

where, by equation (7.13),

$$\mathbb{E}[F_{tr}(i)|\mathbf{y}^{t-1}] = \begin{cases} 1 & \text{if } i = 1, \\ \overline{f_t}(i-1) & \text{if } 2 \leq i \leq r \text{ and } Y_t(i-1) \in Pa(r), \\ 0 & \text{otherwise.} \end{cases} \tag{7.15}$$

The variance of the marginal forecast distribution of $Y_t(r)$ given the past is calculated as:

$$\sigma_t(r,r) = \mathbb{E}\left[\text{var}\{Y_t(r)|\mathbf{y}^{t-1}, \mathbf{F}_t(r)\}|\mathbf{y}^{t-1}\right] + \text{var}\left[\mathbb{E}\{Y_t(r)|\mathbf{y}^{t-1}, \mathbf{F}_t(r)\}|\mathbf{y}^{t-1}\right]$$

$$\begin{aligned} &= \mathbb{E}\left[\frac{d_{t-1}(r)}{n_{t-1}(r)-2}\left(\sum_{j=1}^{r}\sum_{k=1}^{r} R^*_{tr}(j,k)F_{tr}(j)F_{tr}(k) + 1\right)|\mathbf{y}^{t-1}\right] + \text{var}\left[\sum_{i=1}^{r} m^{(i)}_{t-1}(r)F_{tr}(i)|\mathbf{y}^{t-1}\right] \\ &= \frac{d_{t-1}(r)}{n_{t-1}(r)-2}\left(\sum_{j=1}^{r}\sum_{k=1}^{r} R^*_{tr}(j,k)\left[\text{cov}\{F_{tr}(j), F_{tr}(k)|\mathbf{y}^{t-1}\} + \mathbb{E}\{F_{tr}(j)|\mathbf{y}^{t-1}\}\mathbb{E}\{F_{tr}(k)|\mathbf{y}^{t-1}\}\right] + 1\right) + \\ &+ \sum_{i=1}^{r}\sum_{l=1}^{r} m^{(i)}_{t-1}(r)m^{(l)}_{t-1}(r)\text{cov}\{F_{tr}(i), F_{tr}(l)|\mathbf{y}^{t-1}\}, \end{aligned} \tag{7.16}$$

where $\mathbb{E}[F_{tr}(i)|\mathbf{y}^{t-1}]$ is given by equation (7.15) and

$$\text{cov}[F_{tr}(i), F_{tr}(l)|\mathbf{y}^{t-1}] = \begin{cases} \sigma_t(i-1,l) & \text{if both } Y_t(i-1) \text{ and } Y_t(l-1) \text{ belong to } Pa(r) \\ & \text{for } i > 1 \text{ and } l > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now, to completely specify the forecast covariance matrix $\mathbf{\Sigma}_t(r)$, it is necessary to find the marginal covariance between $Y_t(r)$ and $\mathbf{X}_t(r)$, *i.e.*,

$$\boldsymbol{\sigma}_t(r) = (\sigma_t(r,1), \ldots, \sigma_t(r,r-1))' = \text{cov}[Y_t(r), \mathbf{X}_t(r)|\mathbf{y}^{t-1}].$$

To simplify the calculations, we will use this following result considering two variables, say H and Z,

$$\mathbb{E}[H,Z] = \mathbb{E}[\mathbb{E}\{HZ|H\}] = \mathbb{E}[H\mathbb{E}\{Z|H\}].$$

Suppose now that $H = \mathbf{X}_t(r)$ and $Z = Y_t(r)$ and then,

$$\begin{aligned}
\boldsymbol{\sigma}_t(r) &= \mathbb{E}[\mathbf{X}_t(r)\mathbb{E}\{Y_t(r)|\mathbf{y}^{t-1}, \mathbf{X}_t(r)\}|\mathbf{y}^{t-1}] - \mathbb{E}\{Y_t(r)|\mathbf{y}^{t-1}\}\mathbb{E}\{\mathbf{X}_t(r)|\mathbf{y}^{t-1}\} \\
&= \mathbb{E}[\mathbf{X}_t(r)\sum_{i=1}^{r} m_{t-1}^{(i)}(r)F_{tr}(i)|\mathbf{y}^{t-1}] - \overline{f}_t(r)\overline{\mathbf{f}}_t^{r-1} \\
&= \sum_{i=1}^{r} m_{t-1}^{(i)}(r)\mathbb{E}[\mathbf{X}_t(r)F_{tr}(i)|\mathbf{y}^{t-1}] - \overline{f}_t(r)\overline{\mathbf{f}}_t^{r-1}.
\end{aligned}$$

Writing in detail the expression above, for $l = 1, \ldots, r-1$, and using equation (7.14)

$$\begin{aligned}
\sigma_t(r,l) &= \sum_{i=1}^{r} m_{t-1}^{(i)}(r)\mathbb{E}\{Y_t(l)F_{tr}(i)|\mathbf{y}^{t-1}\} - \left[\sum_{i=1}^{r} m_{t-1}^{(i)}(r)\mathbb{E}\{F_{tr}(i)|\mathbf{y}^{t-1}\}\right]\overline{f}_t(l) \\
&= \sum_{i=1}^{r} m_{t-1}^{(i)}(r)\left[\mathbb{E}\{Y_t(l)F_{tr}(i)|\mathbf{y}^{t-1}\} - \mathbb{E}\{F_{tr}(i)|\mathbf{y}^{t-1}\}\overline{f}_t(l)\right] \qquad (7.17) \\
&= \sum_{i=1}^{r} m_{t-1}^{(i)}(r)\mathrm{cov}[Y_t(l), F_{tr}(i)|\mathbf{y}^{t-1}],
\end{aligned}$$

where

$$\mathrm{cov}[Y_t(l), F_{tr}(i)|\mathbf{y}^{t-1}] = \begin{cases} \sigma_t(i-1,l) & \text{if } 2 \leq i \leq r \text{ and } Y_t(i-1) \in Pa(r) \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the marginal forecast covariance between $Y_t(r)$ and $Y_t(l)$ is found by the covariance between $Y_t(l)$ and the parents of $Y_t(r)$. Note that this marginal forecast covariance is zero when the nodes $Y_t(r)$ and $Y_t(l)$ do not have parents. Finally,

$$\overline{\mathbf{f}}_t^r = (\overline{\mathbf{f}}_t^{r-1'}, \overline{f_t}(r))'$$

and

$$\boldsymbol{\Sigma}_t(r) = \begin{pmatrix} \boldsymbol{\Sigma_t(r-1)} & \boldsymbol{\sigma}_t(r) \\ \boldsymbol{\sigma}_t(r)' & \sigma_t(r,r) \end{pmatrix}.$$

For $r = 1$, $\overline{\mathbf{f}}_t^r = \overline{f}_t(1) = m_{t-1}^{(1)}(1)$ and $\boldsymbol{\Sigma}_t(r) = \sigma_t(1,1) = \frac{d_{t-1}(1)}{n_{t-1}(1)-2}[R_t^*(1)+1]$. Then, for $r > 1$, $\overline{\mathbf{f}}_t^r$ and $\boldsymbol{\Sigma}_t(r)$ can be found through the updated distributions for each conditional component DLMs (equation (3.4)) and the expectation and covariance matrix of the marginal forecast distribution for $\mathbf{X}_t(r)$ (equation (7.14), (7.16) and (7.17)). As said in Chapter 3, Queen *et al.* (2008) calculated the forecast covariance matrix for LMDM as showed

here. In addition, they have found the covariance between model regression components, *i.e.* $\text{cov}[F_{tr}(i)\theta_t^{(i)}(r),\, F_{tl}(j)\theta_t^{(j)}(l)|\mathbf{y}^{t-1}]$ for any i^{th} and j^{th} element of the regression of time series $Y_t(r)$ and $Y_t(l)$, respectively.

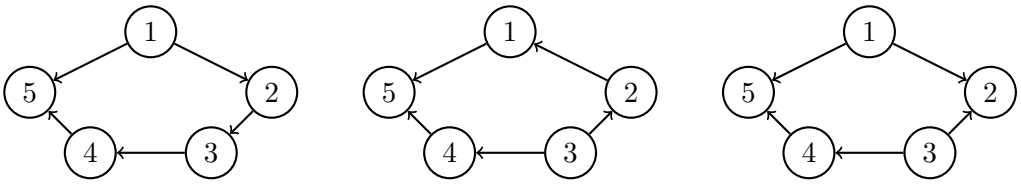
B Supplemental Material for Chapter 4

Appendix B.1: Comparing Markov equivalent DAGs

Here we show the reproducibility of the analysis provided in Section 4.1, now using the synthetic data from the DCM fMRI forward model (Smith, S.M. *et al.*, 2011). The inference process was led considering the true graphical structure, DAG6 (Figure B1 (a)), a Markov equivalent graph, DAG7 (Figure B1 (b)) and a Markov non-equivalent graph, DAG8 (Figure B1 (c)). Considering weakly informative priors, Figure B2 shows the average of LPL across 50 replications for different values of δ and every defined DAG. As a result, the process was estimated correctly as dynamic as long as the chosen δ was different from 1, condition for a static model. Actually, the average of estimated discount factor across nodes and replications was 0.82, 0.86 and 0.84 for DAG6, DAG7 and DAG8 respectively. In addition, on average, DAG6 should be chosen for all values of the discount factor, except for $\delta = 1$ when the LPL is approximately the same for Markov equivalent graphs DAG6 and DAG7, as expected. Finally, the DAG6 was chosen correctly for 88% of replications. Thus, the MDM shows a high success rate considering the synthetic data come from other model, DCM, which defines the graphical structure by the relation between latent variables.

Appendix B.2: Assessment the directionality using the logBF

Here we describe the method used to assess the directionality of the edge, considering the logBF, as cited in Section 4.2.1. That is, for each replication, we fitted the model 0 with the



(a) DAG6 (b) DAG7 (c) DAG8

Figure B1: (a) DAG6: The graphical structure used by Smith, S. M. *et al.* (2011) to simulate data. (b) DAG7 is Markov equivalent to DAG6 whilst neither is equivalent to (c) DAG8. The difference amongst these DAGs is in the relationship between the nodes 1, 2 and 3.

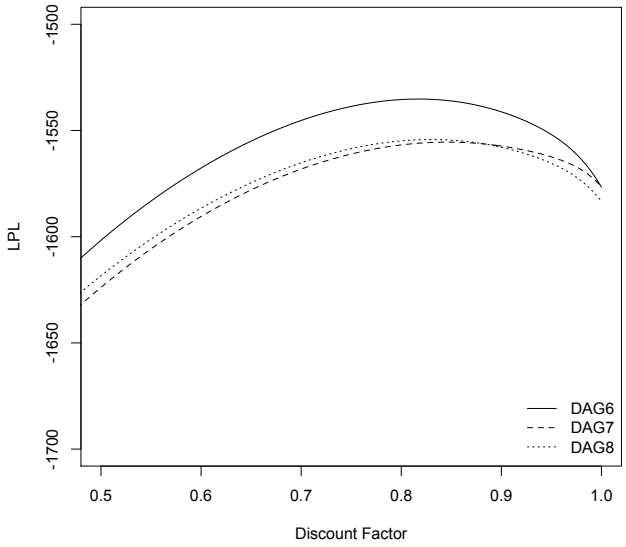


Figure B2: The log predictive likelihood by discount factor using the *sim22* data from Smith, S. M. *et al.* (2011), considering the graphical structures shown in Figure B1, *i.e.* DAG6 (solid line), DAG7 (dashed line) and DAG8 (dotted line).

true connection: node $i \rightarrow$ node j as:

$$\begin{aligned} Y_t(i) &= \theta_t^{(1)}(i) + v_t(i); \\ Y_t(j) &= \theta_t^{(1)}(j) + \theta_t^{(2)}(j) Y_t(i) + v_t(j), \end{aligned} \tag{7.18}$$

for $(i, j) \subset \{(1,2); (2,3); (3,4)\}$. Then we fitted the model 1 with the reverse connection: node $j \rightarrow$ node i, replacing the node i by j and vice versa in the equation (7.18). Thus, the logBF is calculated as:

$$\begin{aligned} logBF &= LPL(\text{model 0}) - LPL(\text{model 1}) \\ &= [LPL(\mathbf{Y}(i)) + LPL(\mathbf{Y}(j)|\mathbf{Y}(i))] - [LPL(\mathbf{Y}(i)|\mathbf{Y}(j)) + LPL(\mathbf{Y}(j)]. \end{aligned}$$

For the collider: node 1 $\rightarrow$ node 5 $\leftarrow$ node 4, the model 0 was fitted as

$$\begin{aligned} Y_t(1) &= \theta_t^{(1)}(1) + v_t(1); \\ Y_t(4) &= \theta_t^{(1)}(4) + v_t(4); \\ Y_t(5) &= \theta_t^{(1)}(5) + \theta_t^{(2)}(5)Y_t(1) + \theta_t^{(2)}(5)Y_t(4) + v_t(5), \end{aligned}$$

whilst the model 1 was considered with reverse edges as node 1 $\leftarrow$ node 5 $\rightarrow$ node 4, and so the observation equations are

$$\begin{aligned} Y_t(1) &= \theta_t^{(1)}(1) + \theta_t^{(2)}(1)Y_t(5) + v_t(1); \\ Y_t(4) &= \theta_t^{(1)}(4) + \theta_t^{(2)}(4)Y_t(5) + v_t(4); \\ Y_t(5) &= \theta_t^{(1)}(5) + v_t(5). \end{aligned}$$

The LPL for a particular model is the sum of LPL for nodes 1, 4 and 5. The proportion of connections selected correctly (logBF>0) is around 70% over all comparisons and all replications. Moreover, the percentage of time that there is evidence to correct model (logBF>1) is 62% over all comparisons and all replications (see Figure 4.8 (right)).

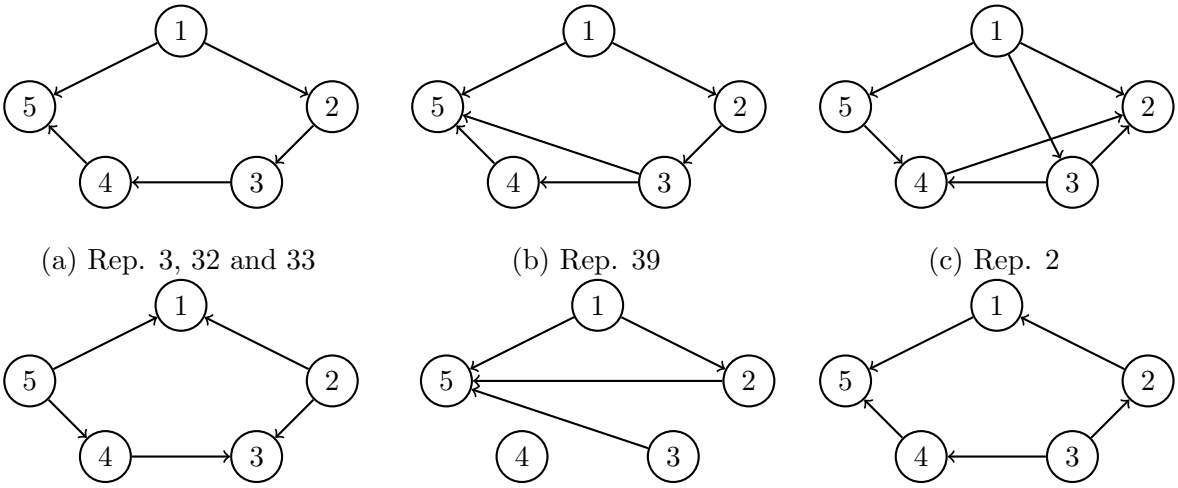
Some estimated DAGs can be seen in Figure B3. The replications with number 3, 32, 33 and 39 had the estimated DAG the closest to true DAG (in fact the graphical structure of three former replications is exactly the same as the true graph), according to logBF comparing the estimated with the true DAG (Figure B3 (a) and (b)). On the other hand, the replications which had the worst results of the learning network process were 2, 34, 44 and 46 (Figure B3 from (c) to (f)).

In addition, the distance between estimated and true structures per node was calculated as the number of parents of a particular node that exist in the estimated DAG but not in true DAG (*false positive parents*) plus the number of parents that exist in the true DAG but not in the estimated DAG (*false negative parents*). In the mathematical language,

$$dist_{T;E}(r) = |Pa_T(r)| + |Pa_E(r)| - |Pa_T(r) \cap Pa_E(r)|,$$

where $dist_{T;E}(r)$ is the distance between the graph T and E with respect to the node r and $|Pa_i(r)|$ is the number of parents of the node r considering the graph i. Figure B4 shows the

relative frequency of the distance over all replications per node, and a general result over all nodes and all replications. To sum up, the distance 0 was predominant in results.



(a) Rep. 3, 32 and 33 (b) Rep. 39 (c) Rep. 2

(d) Rep. 34 (e) Rep. 44 (f) Rep. 46

Figure B3: The estimated DAGs for replications 3, 32, 33 (a) and 39 (b) are the closest to true DAG, with logBF being 0 — it is the true DAG — for replications in (a) and −0.42 for replication 39, whilst the estimated DAGs for replications 2 (c), 34 (d), 44 (e) and 46 (f) are the farthest to the true DAG, with the logBF being −24.79, −41.12, −27.35 and −23.35 respectively.

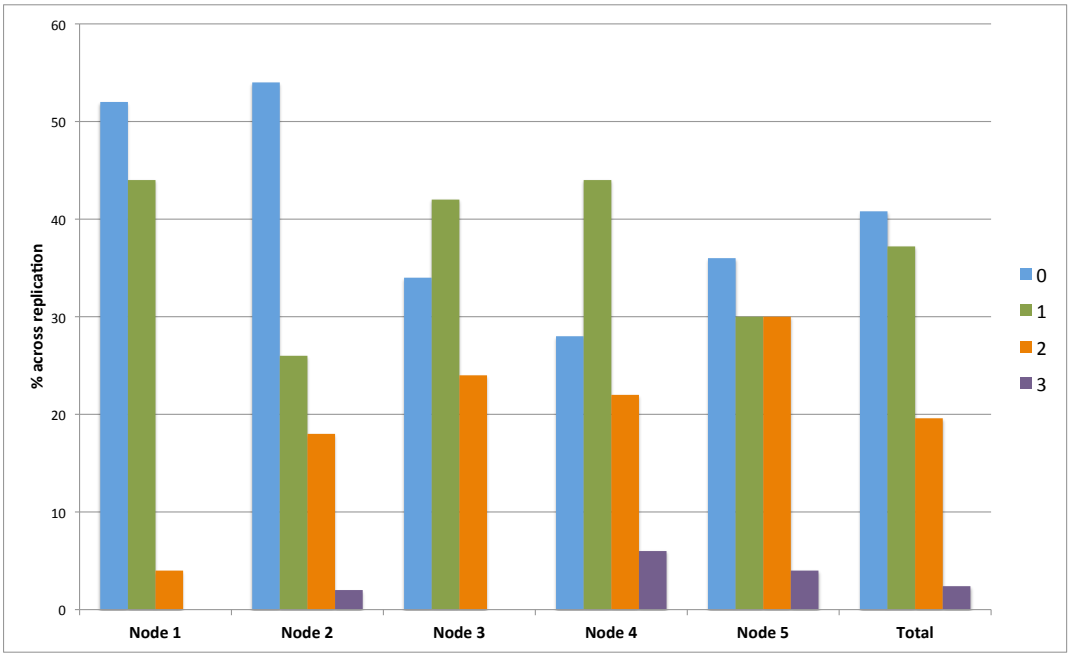


Figure B4: The relative frequency of the distance between the true DAG and the estimated DAG across 50 replications per node and for all nodes (total). The distance is defined as the number of false positive parents plus the number of false negative parents.

C Supplemental Material for Chapter 5

Dynamic Hybrid Algorithm

Dynamic Hybrid Algorithm aims to cluster subjects using information of dendrogram, in a bottom-top manner, and distance measures, in two steps (Langfelder *et al.*, 2008). In the first step, clusters are built according to the following four criteria:

1. There are at least N_0 subjects in each cluster;

2. The *joining heights* of clusters are less than or equal to h_{max}. The joining height is the value of height where a particular cluster joins to the rest of the dendrogram;

3. The average of all pairwise distance between subjects who belong to the same *core* ($\overline{d}_c$) is at most d_{max}. Core consists of the first n_c subjects merged into a cluster, where $n_c = \min\{\text{int}(N_0/2 + \sqrt{N_g - N_0/2}), N_g\}$, and N_g is the total number of subjects in the cluster;

4. The *gap* g is greater than g_{min}. The gap g is the difference between $\overline{d}_c$ and the joining height.

Some terms defined above are illustrated in Figure C1. This algorithm provides flexibility in building cluster, letting users to specify the parameters N_0, h_{max}, d_{max} and g_{mim}. However, Langfelder *et al.* (2008) suggested default values for these parameters, except for minimum cluster size (N_0), which were implemented in R package *dynamicTreeCut*[1].

Subjects who do not belong to any cluster detected in the first step are labelled as *unassigned subjects*. Then, in step 2, the unassigned subjects are included in the closest

[1] *http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/BranchCutting/*

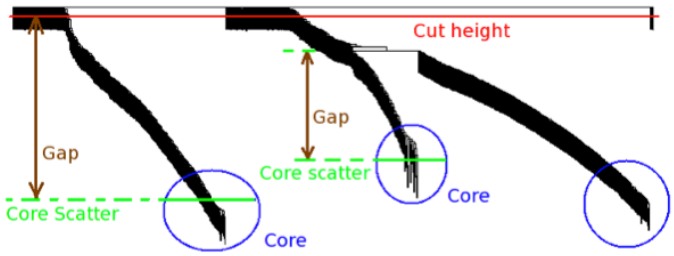


Figure C1: "Illustration of the notions used to define clusters in step 1 of the Dynamic Hybrid algorithm. Shown is a simple simulated dendrogram with 3 branches (clusters) whose joining heights differ. CutHeight corresponds to hmax." (Langfelder *et al.*, 2008, supplementary material).

cluster based on the distance measures as follows:

1. $\overline{d}_{ij}$ is calculated as the average distance between subject i and all other subjects belonging to the same cluster j;

2. The cluster *radius* is defined as the maximum of the average distance of subjects belonging the same cluster, *i.e.* $radius_j = \max(\overline{d}_{1j}, \ldots, \overline{d}_{N_j j})$, where N_j is the number of subjects in the cluster j;

3. $\overline{d}_{u_i j}$ is calculated as the average distance between the unassigned subject u_i and all subjects belonging to the same cluster j;

4. The unassigned subject u_i is included to the closest cluster j if the difference between $\overline{d}_{u_i j}$ and $radius_j$ is the smallest considering all other cluster that satisfy the criteria $\overline{d}_{u_i j} < radius_j$;

5. An unassigned subject who is not included in any cluster is considered outlier.

Multidimensional scaling

The problem of the classical MDS can be written as to transform the original distance $S \times S$ matrix $\mathbf{D}$ into the $q \times S$ matrix $\mathbf{C}$ whose elements represent the coordinates of the $q-$dimensional space of the MDS plot. Here we provide the solution given by Everitt and Hothorn (2009, Chapter 17).

Firstly suppose the $S \times S$ inner products matrix $\mathbf{B} = \mathbf{C}\mathbf{C}'$, whose elements can be

calculated as a function of matrix $\mathbf{D}$, as follows

$$b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - S^{-1}\sum_{j=1}^{S} d_{ij}^2 - S^{-1}\sum_{i=1}^{S} d_{ij}^2 + S^{-2}\sum_{i=1}^{S}\sum_{j=1}^{S} d_{ij}^2\right),$$

where b_{ij} and d_{ij} are the elements at the ith row and jth column of matrix $\mathbf{B}$ and $\mathbf{D}$, respectively. In addition, using singular value decomposition, $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$, where $\mathbf{U}$ is the normalised matrix of eigenvectors so that $\mathbf{U}'\mathbf{U} = \mathbf{I}_S$ whilst $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_S)$ is the matrix of eigenvalues of $\mathbf{B}$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_S$.

The best $q-$dimensional representation consists of the first q eigenvectors and the q largest eigenvalues, so that $\mathbf{B}$ can be approximated by $\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1'$. Thus

$$\mathbf{C} = \mathbf{U}_1\mathbf{\Lambda}_1^{1/2},$$

where $\mathbf{U}_1$ contains the first q eigenvectors and $\mathbf{\Lambda}_1^{1/2} = (\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_q})$.

Mardia *et al.* (1979) suggested the following criterion to assess the adequacy of the $d-$dimensional representation

$$P_q = \frac{\sum_{i=1}^{q}\lambda_i^2}{\sum_{i=1}^{S}\lambda_i^2}.$$

The values of P_q above 0.8 may indicate a good representation.

D Supplemental Material for Chapter 6

Appendix D.1: Supplementary material

The proof of $p(j \notin M_i \Delta \bar{M}_g)$

Here we show more details about the equation (6.2).

$$\begin{aligned}
p(j \notin M_i \Delta \bar{M}_g) &= p(j \in M_i, j \in \bar{M}_g) + p(j \notin M_i, j \notin \bar{M}_g) \\
&= p(j \in M_i | j \in \bar{M}_g) p(j \in \bar{M}_g) + p(j \notin M_i | j \notin \bar{M}_g) p(j \notin \bar{M}_g) \\
&\propto \exp\{-\lambda \times 0\}/2 + \exp\{-\lambda \times 0\}/2 \\
&= \exp\{-\lambda \times 0\}, \text{ and} \\
p(j \in M_i \Delta \bar{M}_g) &= p(j \notin M_i, j \in \bar{M}_g) + p(j \in M_i, j \notin \bar{M}_g) \\
&= p(j \notin M_i | j \in \bar{M}_g) p(j \in \bar{M}_g) + p(j \in M_i | j \notin \bar{M}_g) p(j \notin \bar{M}_g) \\
&\propto \exp\{-\lambda \times 1\}/2 + \exp\{-\lambda \times 1\}/2 \\
&= \exp\{-\lambda \times 1\}.
\end{aligned}$$

Note that we consider here a weak prior, *i.e.* $p(j \in \bar{M}_g) = p(j \notin \bar{M}_g) = 1/2$. Finally we use the fact that $p(j \notin M_i \Delta \bar{M}_g) + p(j \in M_i \Delta \bar{M}_g) = 1$ to write the middle term of equation (6.2).

The False Discovery Rate

Benjamini and Hochberg (1995) suggested an approach for multiple comparisons: the false discovery rate (FDR). This is defined as the expected proportion of incorrect decisions

among the rejected hypotheses, and the procedure is as follows. The p-values are ordered such that $P_{(1)} \leqslant \ldots \leqslant P_{(m)}$, where $P_{(i)}$ is the p-value of the null hypothesis test $H_{(i)}$, for $i = 1, 2, \ldots, m$. All hypothesis $H_{(i)}$, $i = 1, 2, \ldots, k$, are rejected, being k the largest i so that $P_{(i)} \leqslant \frac{i}{m}\alpha$. This procedure controls the FDR at level α.

Lower Tail Test for Population Proportion

Let I^s_{ijk} an indicator function that assumes value 1 if the edge $i \to j$ exists for subject k and session s, and it assumes value 0, otherwise, for i,$j = 1, \ldots, n$ (number of nodes), $k = 1, \ldots, K(s)$ (number of subjects in the session s) and $s = 1, \ldots, S$ (number of sessions). The probability that the edge $i \to j$ exists in the session s is estimated as the proportion of subjects who have this particular edge, *i.e.*

$$\hat{p}^s_{ij} = \frac{\sum_k I^s_{ijk}}{K(s)}.$$

Suppose the probability that one edge exists in the session s, p^s_0, is defined as

$$p^s_0 = \frac{\sum_{ijk} I^s_{ijk}}{K(s)n(n-1)}.$$

To test the null hypothesis $H_0 : p^s_{ij} = p^s_0$ against the alternative hypothesis $H_1 : p^s_{ij} > p^s_0$, the test statistic z^s_{ij} is found as

$$z^s_{ij} = \frac{\hat{p}^s_{ij} - p^s_0}{\sqrt{p^s_0(1-p^s_0)/K(s)}}.$$

Let z_α the $100(1-\alpha)$ percentile of the standard Gaussian distribution, then the null hypothesis is to be rejected if $z^s_{ij} \leqslant -z_\alpha$ (Agresti, 2002, Chapter 1).

McNemar Test for Paired Proportions

Suppose we want to test whether the probability a particular edge $i \to j$ exists in the session s is the same as in the session l. Formally, the null hypothesis $H_0 : p^s_{ij} = p^l_{ij}$ is tested against $H_1 : p^s_{ij} \neq p^l_{ij}$. As the subjects are the same for all sessions, we will show the McNemar test used to compare paired proportions (Agresti, 2002, Chapter 10). The test

statistic is

$$z_{ij}^{sl} = \frac{n_{ij}^{s_1 l_0} - n_{ij}^{s_0 l_1}}{\sqrt{n_{ij}^{s_1 l_0} + n_{ij}^{s_0 l_1}}},$$

where $n_{ij}^{s_1 l_0}$ is the number of subjects who have the edge $i \to j$ in session s but do not have in session l. In contrast, $n_{ij}^{s_0 l_1}$ is the number of subjects who do not have the edge $i \to j$ in session s but have in session l. Under the null hypothesis, the square of z_{ij}^{sl} follows chi-squared distribution with 1 degree of freedom.

Hypothesis Test for Correlation

Let ρ_{ij}^s the full/partial correlation between the time series of node i and j for session s. The null hypothesis $H_0 : \rho_{ij}^s = \rho_0$ may be tested using the following Fisher transformation:

$$\begin{aligned} Z_{ij}^s &= [F(\rho_{ij}^s) - F(\rho_0)]\sqrt{T-3}, \text{ where} \\ F(\rho) &= \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right), \end{aligned}$$

T is the sample size and then the null hypothesis cited above may be written in function of Z_{ij}^s (Hotelling, 1953). Thus, suppose we want to test whether the correlation is zero or positive, then we can write the null hypothesis in terms of the expected value of Z as $H_0 : \mathbb{E}[Z_{ij}^s] = 0$ against $H_1 : \mathbb{E}[Z_{ij}^s] > 0$. Considering z_{ijk}^s as the observed value of Z_{ij}^s for subject k, the test statistic is found using the random sample $\mathbf{z}_{ij}^s = \{z_{ij1}^s, \ldots, z_{ijK(s)}^s\}$ as follows.

$$t_{ij}^s = \frac{\overline{z}_{ij}^s}{sd(\mathbf{z}_{ij}^s)/\sqrt{K(s)}},$$

where $\overline{z}_{ij}^s$ and $sd(\mathbf{z}_{ij}^s)$ are mean and standard deviation of $\mathbf{z}_{ij}^s$, respectively, and t_{ij}^s under null hypothesis follows a Student's t-distribution with $K(s)$ degrees of freedom.

Suppose now we want to test the hypothesis that the correlation between two particular nodes is the same in two sessions. In other words, the null hypothesis is $H_0 : \mathbb{E}[Z_{ij}^s] = \mathbb{E}[Z_{ij}^l]$. Recall that the subjects are the same for all sessions, and then we will use the paired t-test (Snedecor and Cochran, 1989). Let $d_{ijk}^{sl} = z_{ijk}^s - z_{ijk}^l$, $\mathbf{d}_{ij}^{sl} = \{d_{ij1}^{sl}, \ldots, d_{ijK(s)}^{sl}\}$, of course $K(s) = K(l)$, and $\overline{d}_{ij}^{sl}$ and $sd(\mathbf{d}_{ij}^{sl})$ are mean and standard deviation of $\mathbf{d}_{ij}^{sl}$, respectively, then

$$t_{ij}^{sl} = \frac{\overline{d}_{ij}^{sl}}{sd(\mathbf{d}_{ij}^{sl})/\sqrt{K(s)}},$$

where, under null hypothesis, t_{ij}^{sl} follows a Student's t-distribution with $K(s)$ degrees of freedom.

Appendix D.2: The use of diagnostics in a high-dimensional fMRI data

Here we illustrate the use of the parent-child monitor and node monitor, using the fMRI data described in Section 6.3. We selected the resting-state session and the subject 7 because its MDM-IPA result contains all edges that are significant in the group analysis (Figure D1), and so in this sense it was a typical experimental subject. Figure D2 (left) provides the smoothed posterior mean for all connectivities that exist in the graph of subject 7 over time whilst Figure D2 (right) shows the discount factor found for every node. Note that the founder nodes in this individual graph — regions 3 and 10 — have the smallest values of δ. As said before, this is scientifically plausible, as a region not driven by external stimuli may indeed be expected to have the noisiest signal.

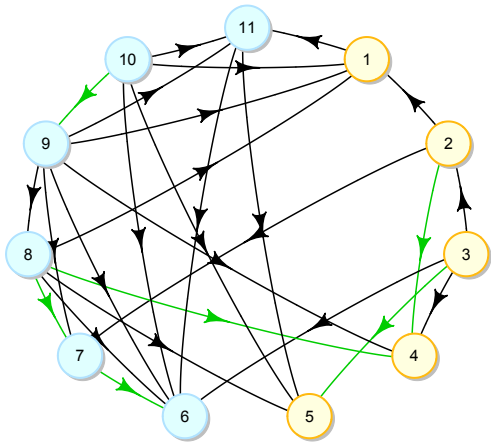


Figure D1: The graphical structure estimated for subject 7 using the MDM-IPA. The green edges are the significant connectivities found in the group analysis (see Figure 6.3 (a)).

We can diagnose and confirm the "parent-child" relationships for the Region 1, as the connectivity from Region 8 into 1 appears to be near zero part of the time. The significance of this connectivity is reflected in

$$\log(BF)_{12} \quad = \quad \log p(\mathbf{y}(1)|\mathbf{y}(2),\mathbf{y}(8),\mathbf{y}(9),\mathbf{y}(10)) - \log p(\mathbf{y}(1)|\mathbf{y}(2),\mathbf{y}(9),\mathbf{y}(10)).$$

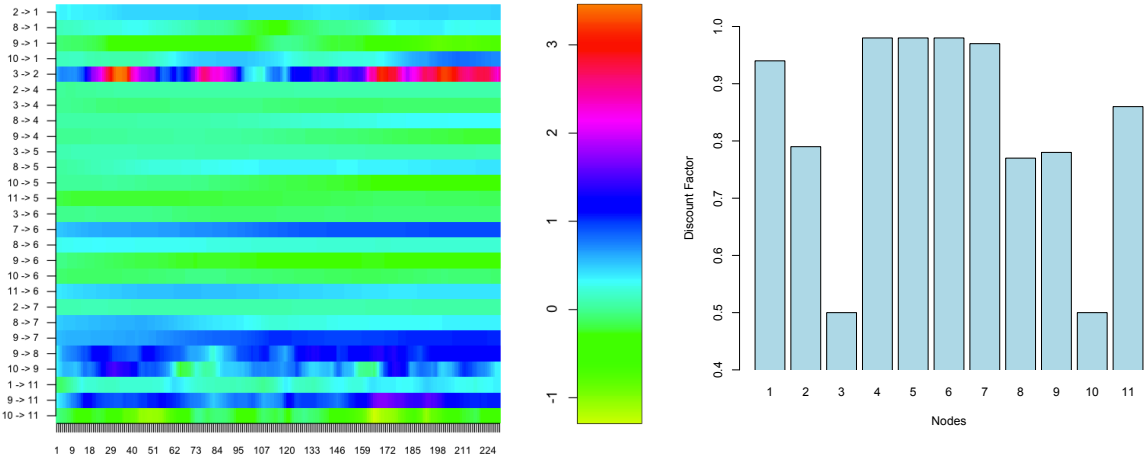


Figure D2: *(Left)* The smoothing posterior mean of connectivities (in *y-axis*) over time (in *x-axis*). *(Right)* Discount factor for each node.

Figure D3 shows the individual contributions to the logBF as well as the cumulative logBF. While the cumulative logBF is close to zero, near time point 110 and again after 180 there is a surge in evidence for the largest model (that includes 8 as a parent).

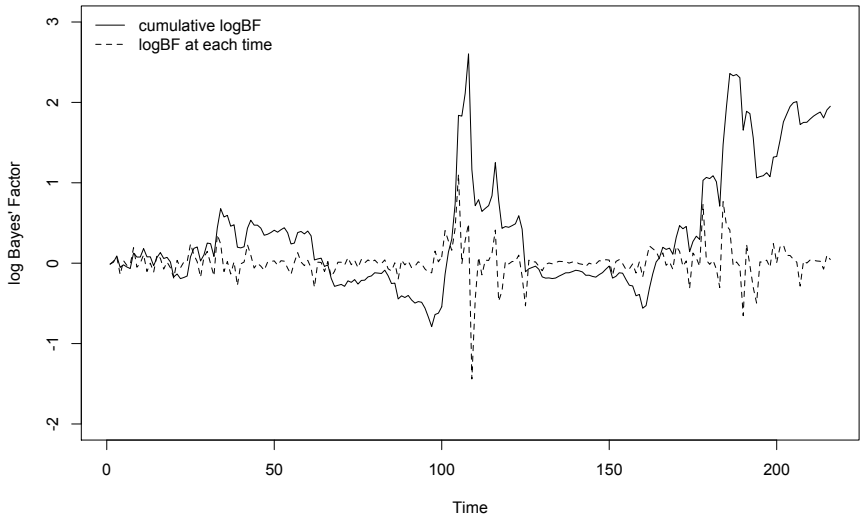


Figure D3: The logBF at each time (dashed lines) and the cumulative logBF (solid lines) comparing Regions 2, 8, 9 and 10 as the set of parents of Region 1 with the same set of parents but without Region 8. The final logBF was 1.95 and, therefore, there is evidence for the former model. This illustrates the use of the parent-child monitor.

As discussed before, a simple LMDM can easily be embellished in order to solve problems detected by diagnostic measures. For example, Figure D4 (first column) shows the time series, the ACF and the cumulative sum plot of the standardised conditional one-step

forecast errors for node 1. Note that the ACF-plot suggests autocorrelation at lag 1. This feature can still be modelled within the MDM class by making a local modification. For example, the past of the region 1 may be included in its observation equation. Thus,

$$\begin{aligned} Y_t(1) &= \theta_t^{(1)}(1) + \theta_t^{(2)}(1)Y_t(2) + \theta_t^{(3)}(1)Y_t(8) + \theta_t^{(4)}(1)Y_t(9) + \\ &+ \theta_t^{(5)}(1)Y_t(10) + \theta_t^{(6)}(1)Y_{t-1}(1) + v_t(1). \end{aligned}$$

Figure D4 (second column) provides the residual analysis plots considering the model with the lag *1*. Although this new model improves the ACF-plot, the cumulative sum of forecast errors (second column and third row) exhibits a non-random pattern, suggesting the presence of change points. Comparing the current graph with the graph where there is no parent from Region 1, and with a threshold of 0.3, two time points were suggested as change points. Figure D5 shows these two change points (dashed lines) and the filtered posterior mean for all connectivities for this region 1, considering the both models, without (blue lines) and with (violet lines) change points. This gives us a different and higher scoring model, one whose score can still be calculated in closed form. Normality and heteroscedasticity tests were also employed in this study, but neither detected any significant deviation from the model class.

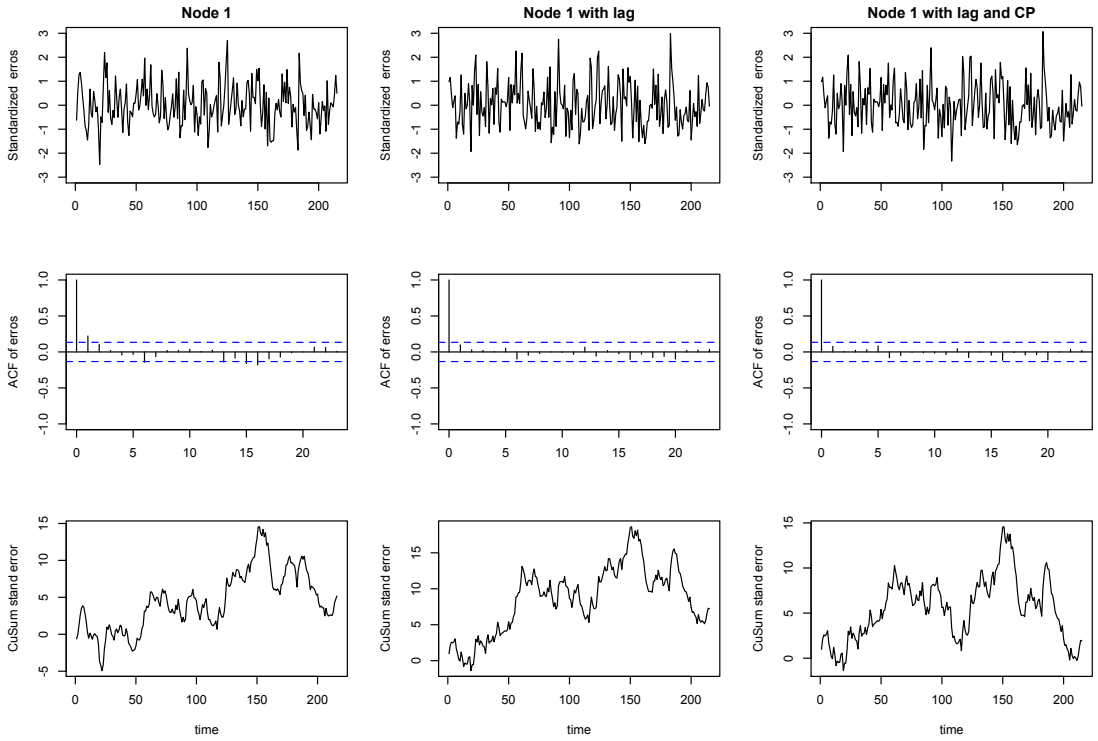


Figure D4: Time series plot, ACF-plot and the cumulative sum of one-step-ahead conditional forecast errors for Region 1 (first column), considering *lag 1* (second column) and considering *lag 1* and *change points* (third column). This illustrates the use of the node monitor.

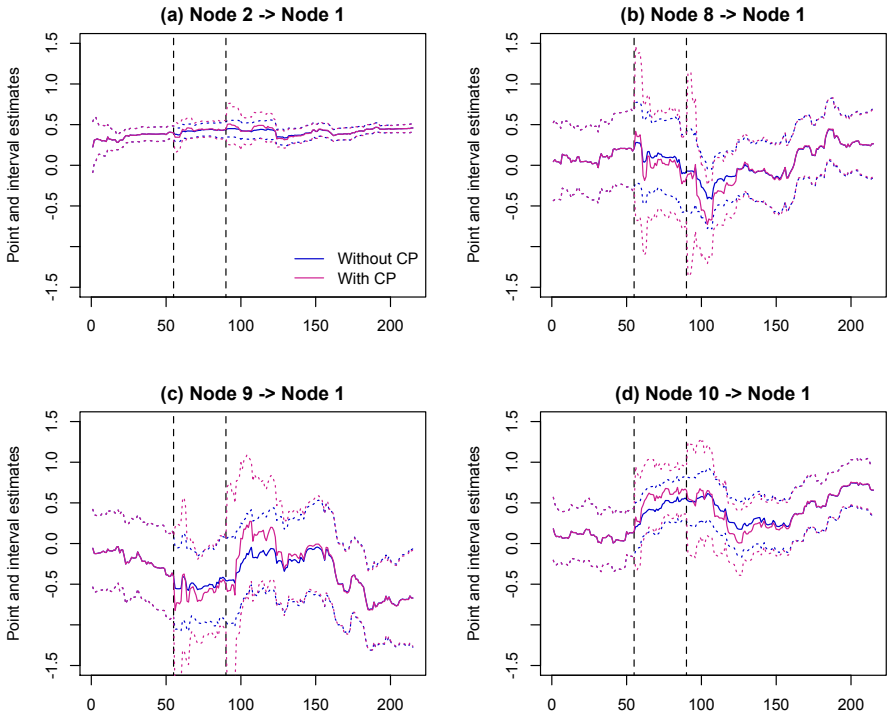


Figure D5: The filtered posterior mean with 95% credible interval for connectivities (a) Region 2 → Region 1, (b) Region 8 → Region 1, (c) Region 9 → Region 1 and (d) Region 10 → Region 1, considering the model without change points (blue lines) and with change points (violet lines). The dashed lines represent the two change points.

Appendix D.3: Additional figures

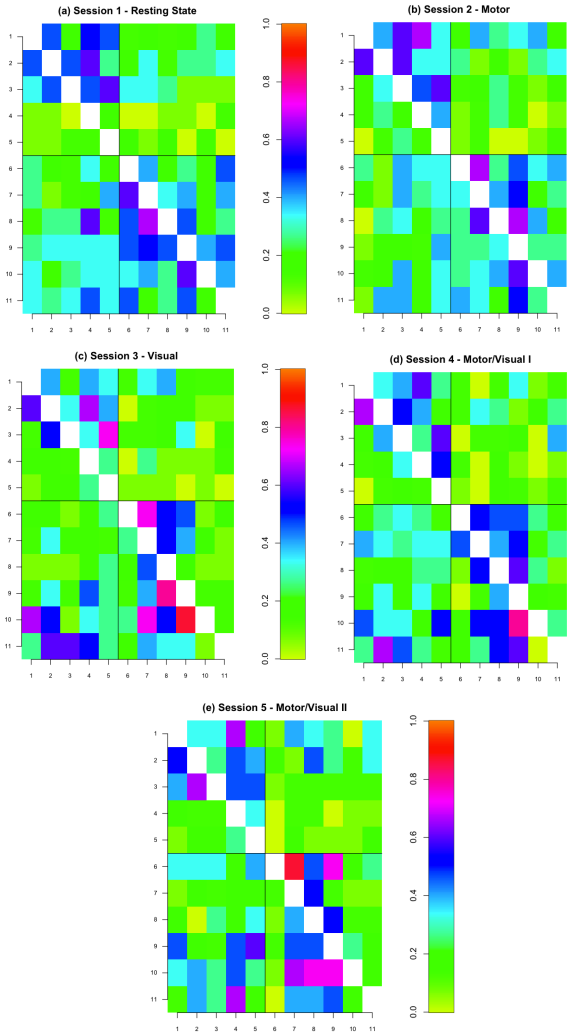


Figure D6: The proportion of subjects who have a particular edge $i \rightarrow j$, where i indexes rows and j columns, using the *MDM-IPA* per session. Nodes numbered from 1 to 5 are motor regions, while nodes numbered from 6 to 11 are visual regions. The black horizontal and vertical lines divide the figure into four squares; the top left square represents the connectivity between motor brain regions, whilst the lower right square represents one between visual brain regions.

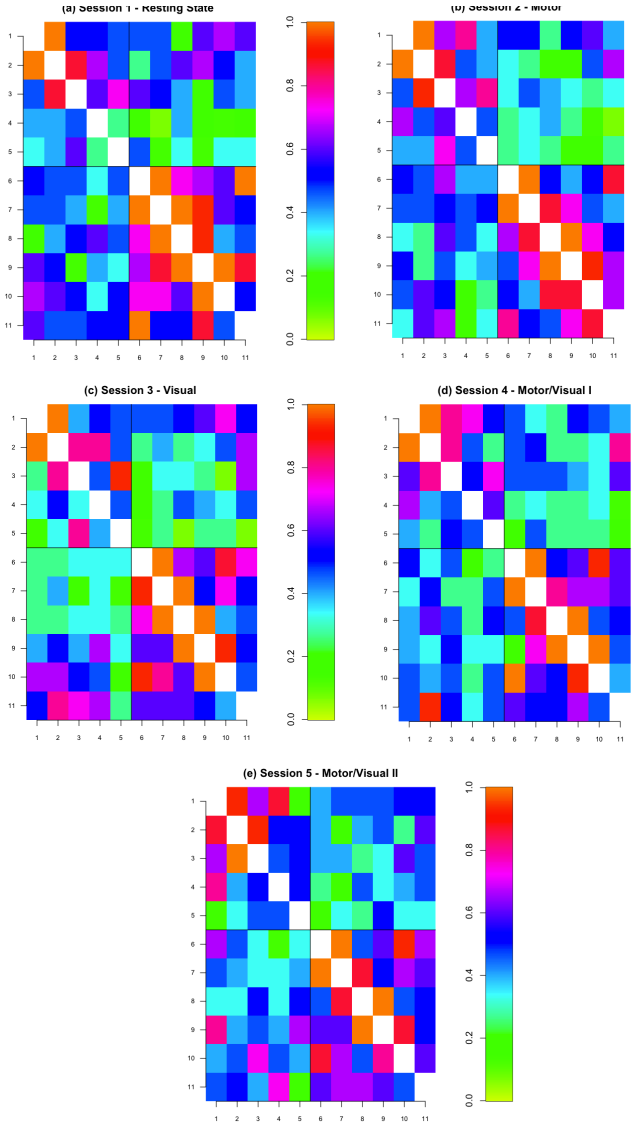


Figure D7: The proportion of subjects who have a particular edge $i \rightarrow j$, where i indexes rows and j columns, using the *MDM-DGM* per session. Nodes numbered from 1 to 5 are motor regions, while nodes numbered from 6 to 11 are visual regions. The black horizontal and vertical lines divide the figure into four squares; the top left square represents the connectivity between motor brain regions, whilst the lower right square represents one between visual brain regions.

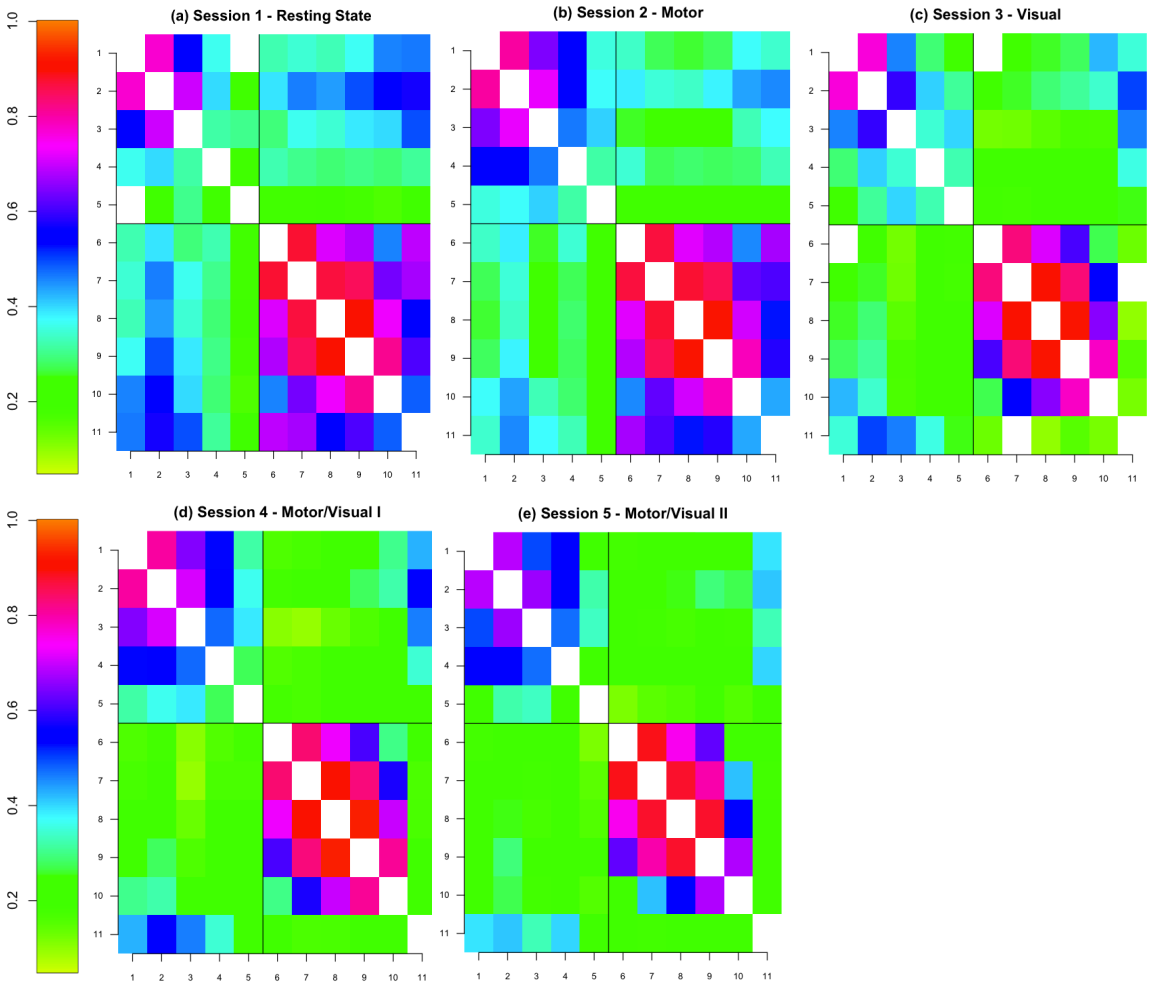


Figure D8: The significant mean *full correlation* between two nodes per session.

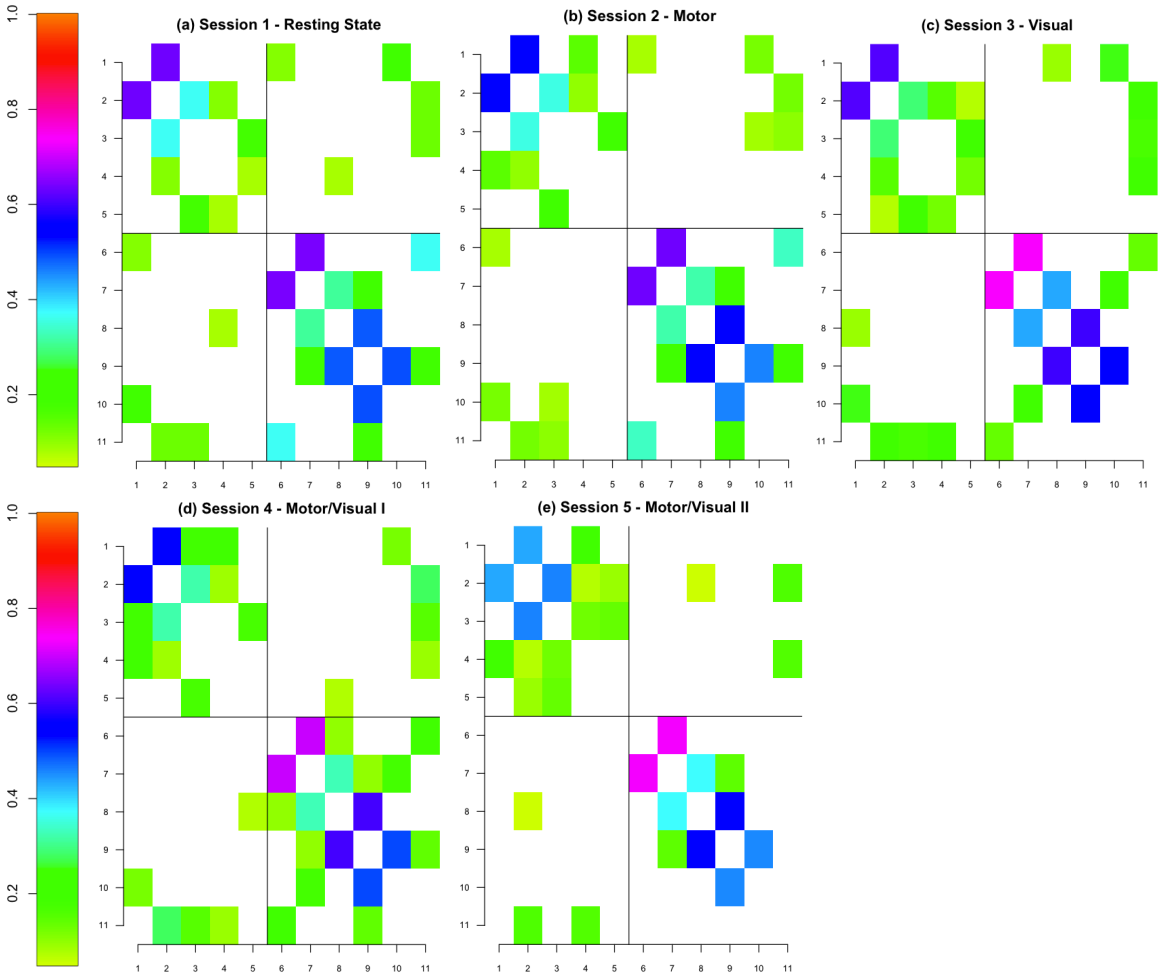


Figure D9: The significant mean *partial correlation* between two nodes per session.

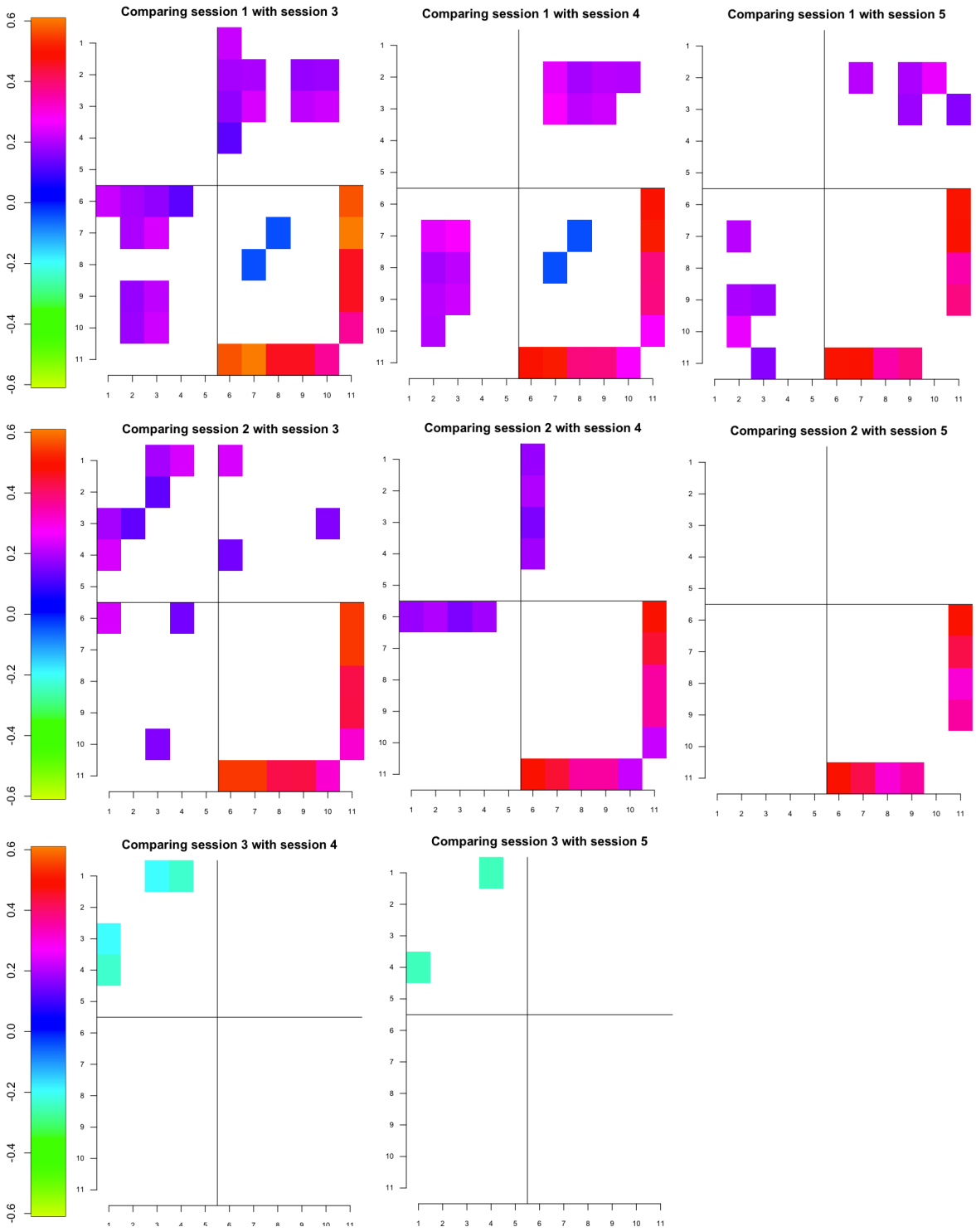


Figure D10: The significant difference of the average of *full correlation* over subjects between two sessions. Session 1 is a resting-state condition; session 2 is a motor condition in which individuals tapped something; session 3 is a visual condition in which individuals watched a movie; session 4 and session 5 are a combination between visual and motor condition, but the former is in a random way whilst in the latter, individuals tapped depending on random events in the movie.

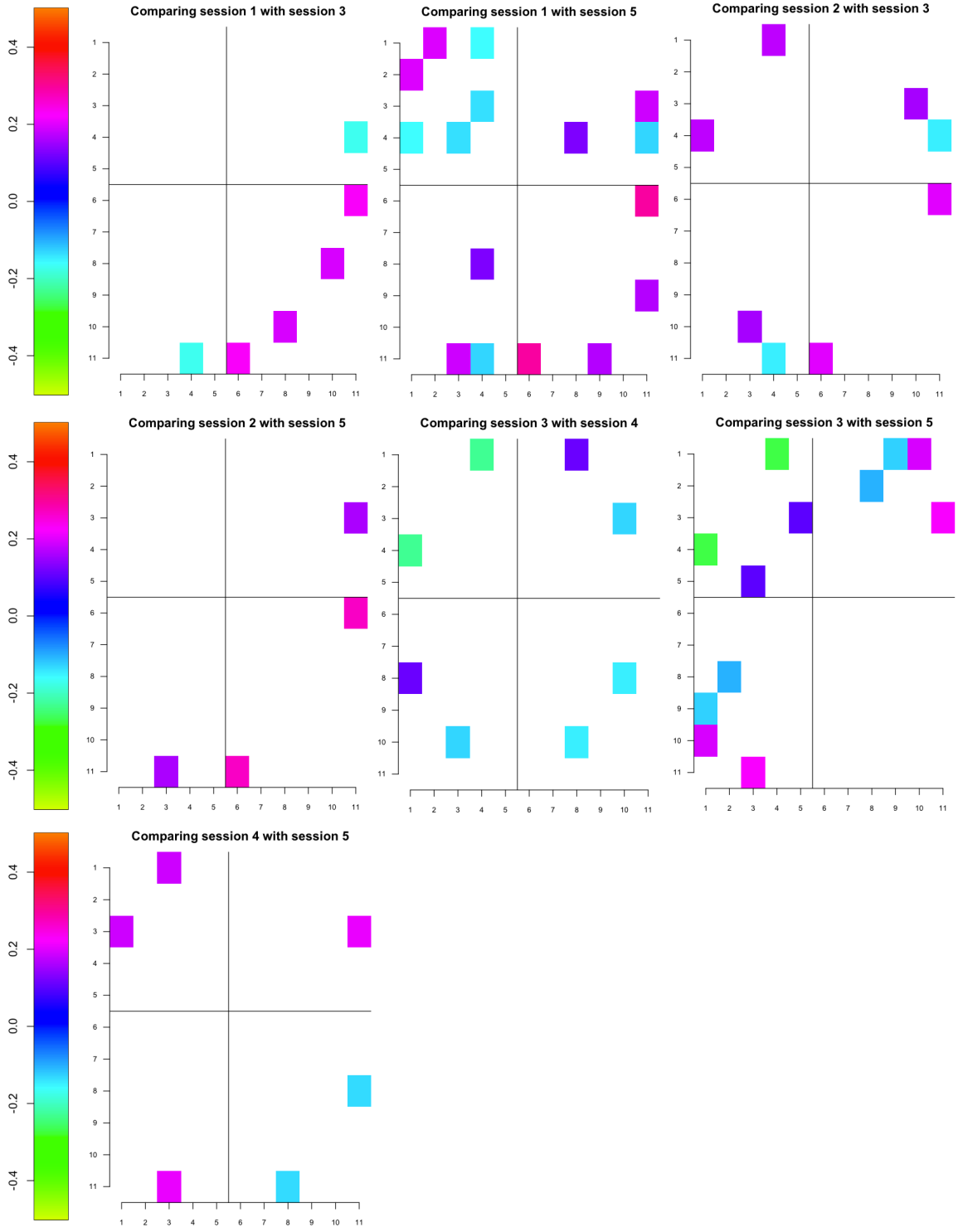


Figure D11: The significant difference of the average of *partial correlation* over subjects between two sessions. Session 1 is a resting-state condition; session 2 is a motor condition in which individuals tapped something; session 3 is a visual condition in which individuals watched a movie; session 4 and session 5 are a combination between visual and motor condition, but the former is in a random way whilst in the latter, individuals tapped depending on random events in the movie.

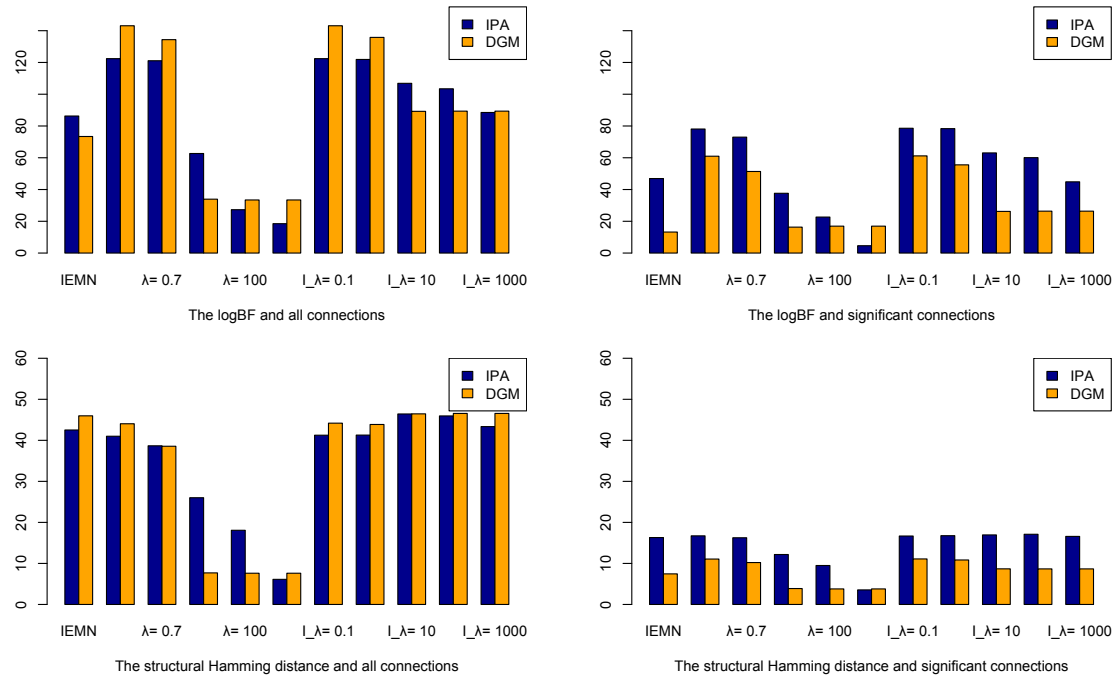


Figure D12: The average of the logBF (top) and the structural Hamming distance (bottom) comparing the estimated to the predicted networks using the same method, *i.e.* the IEMN and the MEMN with $\lambda = 0.1, 0.7, 10, 100, 1000$, and comparing the estimated networks using the IEMN to the predicted networks using the MEMN for $\mathrm{l}\_\lambda = 0.1, 0.7, 10, 100, 1000$, over subjects and sessions, considering the entire graphs (left) and the graphs formed by only significant edges (right), and using the search method the MDM-IPA (blue bars) and the MDM-DGM (orange bars).